UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Classification Methods applied to Familial Hypercholesterolemia Diagnosis in Pediatric Age.

João David Ferreira de Castro Albuquerque

Mestrado em Bioestatística

Dissertação orientada por:
Professora Doutora Marília Antunes
Professora Doutora Mafalda Bourbon

2019

Acknowledgments

Apesar da língua inglesa ter sido a escolhida para a elaboração da presente tese, optei por escrever os respectivos agradecimentos em português, sendo a minha língua materna, bem como a de todos a quem se destinam as próximas linhas.

À minha orientadora, Professora Doutora Marília Antunes, Professora Auxiliar na Faculdade de Ciências da Universidade de Lisboa (FCUL), por todo o apoio, paciência, momentos de ensino, generosidade e motivação. Por ter sido a principal impulsionadora por trás do grande crescimento pessoal e profissional que este mestrado representou.

À minha co-orientadora, Professora Doutora Mafalda Bourbon, coordenadora do Estudo Português de Hipercolesterolemia Familiar (EPHF), no Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA), por me ter integrado no seu grupo de investigação, por ter permitido a utilização de dados do EPHF para este estudo, e pelo seu acompanhamento no desenvolvimento deste trabalho.

A todos os meus colegas investigadores no INSA, em particular à Ana Catarina Alves, à Ana Medeiros, à Joana Chora, à Leonor Abrantes e ao Rafael Graça, por todo o apoio, partilha de ideias e de conhecimento, e momentos de convívio que tornaram os dias de trabalho mais descontraídos e animados.

Aos participantes no EPHF, especialmente os indivíduos em idade pediátrica, que constituíram a amostra do presente estudo. A todos os médicos e outros profissionais de saúde que colaboram com o EPHF, e despendem algum do seu tempo e energia no processo de referenciação e recolha de dados dos participantes;

Aos outros professores do Mestrado em Bioestatística da FCUL, pelos conhecimentos e competências que adquiri ao longo deste período, e que considero serem uma mais-valia para o meu futuro. Aos meus colegas de turma, pelo companheirismo, partilha de conhecimento, e entreajuda. Um agradecimento especial ao João Malato, pelas contribuições para o desenvolvimento desta tese.

À minha família, especialmente aos meus pais, Alfredo e Maria, e ao meu irmão Miguel, pelo apoio, carinho e encorajamento constante, que me permitem encarar cada desafio com motivação e confiança, e por me ajudarem a valorizar e a equilibrar todas as vertentes importantes da vida.

À Mara, por todo o teu amor. Por estares presente em todos os momentos e tornares cada dia mais feliz.

João Albuquerque

Lisboa, Setembro de 2019.

Resumo

Introdução: A Hipercolesterolemia Familiar (FH) é uma doença genética do metabolismo lipídico, caracterizada por níveis elevados de colesterol proveniente das lipoproteínas de baixa densidade (LDLc). A severa dislipidemia resultante leva ao desenvolvimento precoce de aterosclerose, representando um grande factor de risco de doença cardiovascular (CVD). O diagnóstico antecipado da FH encontra-se associado com uma redução significativa do risco de CVD, fundamentando a introdução de medidas terapêuticas mais precoces e agressivas. Existem diferentes critérios clínicos disponíveis para o diagnóstico da FH, sendo que apenas através de teste genético se pode confirmar o mesmo. Os critérios de Simon Broome (SB) para o diagnóstico da FH são dos mais frequentemente utilizados em contexto clínico, e são baseados na história familiar, presença de sinais físicos, e concentração plasmática de LDLc e colesterol total (TC). Quando comparados com os resultados do diagnóstico genético contudo, os critérios de SB apresentam uma elevada taxa de falsos positivos, o que constitui um pesado fardo em termos de despesas de saúde, e limita o acesso ao estudo molecular por parte de um maior universo de potenciais casos de FH.

Objectivos: O objectivo principal do presente estudo foi desenvolver métodos de classificação alternativos para o diagnóstico da FH, a partir de diferentes indicadores bioquímicos, que pudessem demonstrar melhor capacidade para rastrear esta patologia comparativamente aos critérios de SB. Dois modelos distintos foram desenvolvidos para este propósito: um modelo de regressão logística (LR) e um modelo em árvore de decisão (DT).

Métodos: Concentrações séricas de TC, LDLc, colesterol associado às lipoproteínas de alta densidade (HDLc), triglicerídeos (TG), apolipoproteinas AI (apoAI) e B (apoB), e lipoproteína(a) (Lp(a)) foram determinadas, e o diagnóstico molecular foi efectuado, numa amostra de 252 participantes no estudo Português de FH, em idade pediátrica (2-17 anos). Todos os participantes possuíam os critérios clínicos de dislipidemia, e não se encontravam sob medicação hipolipidémica durante o período de avaliação. Os modelos de LR e DT foram ajustados aos dados da amostra. Para o modelo de LR, dois valores de corte distintos foram definidos, através de análise de curvas ROC (*receiver operating characteristics*), de acordo com os métodos do índice de Youden e mínimo valor-p (*min* p). A construção da DT foi baseada em medidas de redução da entropia, ou ganho de informação. Uma versão modificada da DT foi implementada, na qual se procedeu à exclusão sequencial de variáveis á medida que eram incluídas no modelo. Este processo permite produzir uma regra de classificação que utiliza valores de corte únicos para cada biomarcador, simplificando a sua interpretação. Diferentes características operacionais (OC) foram estimadas para todos os modelos: acurácia (Acc), sensibilidade (Se), especificidade (Spe), valor preditivo positivo (PPV) e valor preditivo negativo (NPV). Estas OC foram calculadas através de uma matriz de confusão, considerando os resultados do teste molecular como o verdadeiro estado da doença. O modelo de LR e a DT com melhor desempenho foram comparados com os critérios bioquímicos de SB, através de técnicas de *bootstrap resampling*. Os valores da média e da mediana para as OC de 200 amostras

bootstrap foram utilizados para comparação da performance preditiva dos modelos.

Resultados: A função *logit* para o modelo de LR final foi expressa como $\widehat{g(\pi)} = -7.083 + 0.086 \times LDLc - 0.041 \times TG - 0.037 \times apoAI$. O modelo DT com melhor desempenho incluiu as variáveis LDLc, TG, apoAI, apoB e HDLc, por ordem descendente de importância. Entre os diferentes métodos de classificação, os valores de *Acc*, *Spe* e PPV foram mais elevados para o modelo DT, seguido do modelo LR com valor de corte (c) definido pelo método *min p* ($c = 0.35$). Os valores mais reduzidos para estas OC são encontrados com os critérios de SB ($p < 0.01$). Valores mais elevados de *Se* e NPV por outro lado, são alcançados pelos critérios de SB, e pelo modelo de LR com o valor de corte calculado através do índice de Youden ($c = 0.17$). O modelo de LR utilizando este ponto de corte revela contudo valores significativamente mais elevados de *Acc*, *Spe* e NPV ($p < 0.01$) em relação aos critérios de SB.

Conclusões: Tanto o modelo de LR como DT parecem ser alternativas válidas aos tradicionais critérios clínicos para diagnóstico da FH. Parece ser possível ajustar o valor de corte do modelo de LR para obter níveis de *Se* similares aos observados para os critérios de SB, com uma retenção de casos falsos positivos significativamente menor. A validação destes resultados por dados adicionais, indicaria indubitavelmente este método como preferível entre os dois, e poderá ter um impacto muito significativo em termos de relação custo-efectividade. Ao evitar a repetição de variáveis predictoras, e providenciar valores de corte únicos para cada biomarcador, o modelo DT modificado assume uma estrutura que se assemelha aos critérios médicos clássicos, e pode portanto ser facilmente utilizado na prática clínica. Parece que, apesar de serem baseados em metodologias distintas, tanto o modelo de LR como a DT são capazes de dividir a amostra de acordo com os indicadores bioquímicos mais relevantes para o diagnóstico da FH. De acordo com ambos os métodos de classificação, a presença de FH encontra-se directamente relacionada com os níveis de LDLc, e inversamente relacionada com as concentrações de TG e apoAI, por esta ordem de importância. O modelo de classificação preferido, assim como as especificações do mesmo, podem variar em função das OC que são consideradas mais importantes, e do contexto em que este é aplicado.

Palavras-chave: Hipercolesterolemia Familiar; Regressão Logística; Árvore de Decisão; critérios de Simon Broome; *Bootstrap Resampling*.

Abstract

Introduction: Familial Hypercholesterolemia (FH) is an inherited disorder of lipid metabolism, characterized by increased low density lipoprotein cholesterol (LDLc) levels. The resulting severe dyslipidemia leads to the early development of atherosclerosis, representing a major risk factor for cardiovascular disease (CVD). The early diagnosis of FH is associated with a significant reduction in CVD risk, supporting the introduction of precocious and more aggressive therapeutic measures. There are different clinical criteria available for the diagnosis of FH, although only genetic testing can confirm the diagnostic. Simon Broome (SB) criteria for FH diagnosis are among the most frequently used in clinical setting, and are based on family history, presence of physical signs, and LDLc and total cholesterol (TC) levels. When compared to genetic diagnosis results however, SB criteria present a high false positive rate, which constitutes a heavy burden in terms of healthcare costs, and limits the access to the genetic study of a larger universe of potential FH cases.

Aim: The main purpose of this work was to develop alternative classification methods for FH diagnosis, based on different biochemical indicators, with improved ability to screen for FH cases in comparison to SB criteria. Two different models were developed for this purpose: a logistic regression (LR), and a decision tree (DT) model.

Methods: Serum concentrations of TC, LDLc, high density lipoprotein cholesterol (HDLc), triglycerides (TG), apolipoproteins AI (apoAI) and B (apoB), and lipoprotein(a) (Lp(a)) were determined, and genetic diagnosis was performed, in a sample of 252 participants in the Portuguese FH Study, at pediatric age (2-17 years). All patients met the clinical criteria for dyslipidemia, and were not under hypolipidemic medication during the evaluation period. LR and DT models were fitted to sample data. For the LR model, two different cutoff points were defined, through receiver operating characteristics (ROC) curve analysis, following Yoden index and minimum p-value ($min\ p$) methods. The DT was built based on entropy reduction, or information gain measures. A modified version of the DT method was implemented, consisting in the sequential exclusion of predictor variables as they are introduced in the model. This allows producing a classification rule that uses single cut-points for biomarkers, simplifying its interpretation. Different operating characteristics (OC) were estimated for all models: accuracy (Acc), sensitivity (Se), specificity (Spe), positive predictive value (PPV) and negative predictive value (NPV). These OC were calculated by generating a confusion matrix, considering molecular study results as the true state of the disease. The best performing LR and DT models were compared with SB biochemical criteria for FH diagnosis, through bootstrap resampling techniques. Median and mean values of the OC for 200 bootstrap samples were used for predictive performance comparison.

Results: The logit function for the LR final model was expressed as $\widehat{g(\pi)} = -7.083 + 0.086 \times LDLc - 0.041 \times TG - 0.037 \times apoAI$. The best performing DT model included the variables LDLc, TG, apoAI, apoB and HDLc, by descending order of importance. Between the different classification methods, Acc, Spe and PPV were higher in the DT model, followed by the LR model with the cut

point value (c) defined by the *min p* method ($c = 0.35$). The lower values in these OC are found for SB criteria ($p < 0.01$). Higher Se and NPV on the other hand, are achieved by SB criteria, and the LR model with the cutpoint value calculated by Youden index ($c = 0.17$). However, the LR model using this cutpoint achieves significantly higher Acc, Spe and NPV than SB criteria ($p < 0.01$).

Conclusions: Both LR and DT models seem to be a valid alternative to traditional clinical criteria for FH diagnosis. It seems possible to adjust the cutoff value in the LR model for similar Se levels as the ones observed in SB criteria, with significantly less false positive retention. To be validated by additional data, this would undoubtedly indicate this method as preferable between the two, and can have a very important impact in terms of cost-effectiveness. By avoiding the repetition of predictor variables, and providing single cutoff values for each biomarker, the modified DT model assumes a structure that typically resembles medical criteria, and can therefore be easily used in clinical practice. It seems that, in spite using different methodological approaches, both LR and DT models are able to divide the sample according to the most relevant biochemical characteristics for FH diagnosis. According to both classification methods, presence of FH is directly related to LDLc levels, and inversely related to TG and ApoAI concentrations, by this order of importance. The preferred classification model, as well as model specifications, may vary as a function of the OC that are considered more important, and context in which it is applied.

Keywords: Familial Hypercholesterolemia; Logistic Regression; Decision Tree; Simon Broome criteria; Bootstrap Resampling.

Contents

List of Figures

List of Tables

List of Abbreviations

FH	Familial Hypercholesterolemia
HDL	High Density Lipoproteins
Lp(a)	Lipoprotein(a)
LDL	Low Density Lipoproteins
IDL	Intermediate Density Lipoproteins
VLDL	Very Low Density Lipoproteins
TC	Total Cholesterol
Apo	Apolipoprotein
LPL	Lipoprotein Lipase
LRP	LDL receptor Related Protein
HepL	Hepatic Lipase
CVD	Cardiovascular Disease
SR-B1	Scavenger Receptor B1
HeFH	Heterozygous FH
HoFH	Homozygous FH
LDLR	Low Density Lipoprotein Receptor gene
APOB	Apolipoprotein B gene
PCSK9	Proprotein Convertase Subtilisin Kexin type 9 gene
CHD	Coronary Heart Disease
SB	Simon Broome
EAS	European Atherosclerosis Society
LR	Logistic Regression
GLM	General Linear Model
IRLS	Iterative Reweighted Least Squares

SE	Standard Error
CI	Confidence Interval
OR	Odds Ratio
AIC	Akaike Information Criteria
OLS	Ordinary Least Squares
GOF	Goodness of Fit
HL	Hosmer-Lemeshow
LC	Le Cessie
OC	Operating Characteristics
Acc	Accuracy
Se	Sensitivity
Spe	Specificity
PPV	Positive Predictive Value
NPV	Negative Predictive Value
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve
DT	Decision Tree
INSA	Instituto Nacional de Saúde Doutor Ricardo Jorge
MLPA	Multiplex Ligation-dependent Probe Amplification
PCR	Polymerase Chain Reaction
KS	Kolmogorov-Smirnov
MWW	Mann-Whitney-Wilcoxon

Chapter 1

Introduction

Familial hypercholesterolemia (FH) is an autosomal dominant disorder of lipid metabolism, characterized by high plasmatic cholesterol concentrations, in particular low density lipoproteins cholesterol (LDLc) [1]. The pathology is caused by mutations in genes that encode essential proteins involved in the LDL receptor (LDLr) metabolic pathway, mainly variants in the LDL receptor (*LDLR*) gene, and less frequently in apolipoprotein B (*APOB*) and proprotein convertase subtilisin kexin type 9 (*PCSK9*) genes [2]. This results in lifelong severely increased cholesterol levels, which lead to the early development of atherosclerosis, and represent a major risk factor for cardiovascular disease (CVD) [1, 3].

The early diagnosis of FH has been associated with a significant reduction in CVD risk, supporting the introduction of precocious and/or more aggressive therapeutic measures, in a cost-effective fashion [4]. There are different clinical criteria available for the diagnosis of FH, although only genetic testing can confirm the diagnosis [5]. Simon Broome (SB) criteria are among the most frequently used in clinical setting, and are based on total cholesterol (TC) and LDLc plasma concentrations, presence of tendinous xanthomas and family history [6]. One of the major problems presented by clinical diagnostic criteria is the high false positive rate they present when compared to molecular study results. [7]. This issue constitutes a heavy burden in terms of healthcare costs, limiting the access to the genetic study of a larger universe of potential FH cases, at an earlier stage.

The main motivation behind the present work was therefore the improvement of the ability to detect potential FH cases, providing an alternative to the currently used SB criteria. For this purpose, two different classification methods have been developed: a logistic regression (LR) model, an approach based on classical inferential methods, and a decision tree (DT) model, used in information theory.

LR is a special case of the generalized linear models (GLM), used to analyse the relationship between a categorical dependent variable, and one or several independent variables. In this case the dependent variable is binary, since the outcome can only assume one of two values, depending on whether the subject is FH or not. The expected value of the dependent variable given by the logistic function ranges between 0 and 1, and represents the probability of this outcome variable to be FH. It is therefore needed to select the cutoff point that best differentiates FH from non FH subjects to use as a classification rule. A receiver operating characteristics (ROC) curve analysis was performed

for this purpose, using two distinct methods to select the optimal cutoff point.

In a DT model, the classification rule is created by repeatedly dividing the data into smaller and increasingly more homogeneous groups, with respect to a certain variable of interest, a method defined as recursive partitioning. One of the biggest advantages of this model is that it visually assumes a tree-like structure, composed of nodes and branches, in which the classification rule can be easily interpreted. Since the target variable is categorical in this case (presence or absence of disease), the DT is named a classification tree. According to this method, data is consecutively divided based on thresholds in the predictor variables, until no further splitting is possible, or additional splits do not improve classification performance. Different algorithms can be used to select the predictor variable and respective cutoff point that best divides the sample at each node, i.e. that most reduces overall node impurity. The method selected for the current study was information gain, or entropy reduction algorithm. To avoid over-fitting, an internal cross-validation is finally performed to remove splits which decrease true classification rate, a process referred to as pruning. Additionally, a modified version of the DT method was developed for this study, consisting in the sequential exclusion of predictor variables as they are introduced in the model. This was done so that each of the predictor variables would enter the DT at most only once, associated with a single cutoff value, assuming a structure that resembles typical medical criteria.

The sample used in the current study was taken from the Portuguese FH study, ongoing since 1999 with the purpose of diagnosing and characterizing FH in the Portuguese population. Data from 252 participants in the Portuguese FH study between 2 and 17 years of age, of both sexes, was used to build LR and DT classification models. A panel of several biochemical variables related to lipid metabolism was used as candidate predictor variables: TC, LDLc, high density lipoprotein cholesterol (HDLc), triglycerides (TG), apolipoproteins AI (apoAI) and B (apoB), and lipoprotein(a) (Lp(a)). Different LR and DT models were developed, using different criteria, and compared through bootstrap resampling techniques. The best performing LR and DT model were finally compared with each other, as well as with SB biochemical criteria for FH diagnosis, using the same bootstrap samples, to allow a global comparison of results. Median and mean values of several operating characteristics (OC) of 200 bootstrap samples were used for performance comparison. For each of the bootstrap samples, respective OC were calculated by generating a confusion matrix for each classification method, considering molecular study results as the reference standard.

The results from this study will hopefully provide new insights on classification methods for FH diagnosis. The preferred method of classification may vary as a function of the clinical decider main goal, as p.e. to minimize overall error of classification, or to maximize true FH cases detection rate.

Chapter 2

Literature Review

This review of the literature is divided in two sections. The first section comprises a brief description of lipoproteins function and structure, as well as the main mechanisms underlying lipoproteins metabolic pathways. This overview of biomarkers function and main interactions with other molecules is thought to be important, in order to understand the biological relevance of the variables used in the study, besides what the statistical significance may prove to be. In the second section, several important concepts related to familial hypercholesterolemia (FH) definition and management are addressed, namely epidemiological data, the molecular characterization of FH, current clinical criteria used for diagnosis, forms of treatment available, and a presentation of the Portuguese FH study, from which the sample of this work was withdrawn.

2.1 Lipoproteins Metabolism

2.1.1 Lipoproteins function and structure

Due to their insoluble properties, in blood plasma, lipids are predominantly transported in sphere like structures, called lipoproteins. Plasmatic lipoproteins are constituted by a central hydrophobic core of non-polar lipids, essentially triglycerides and cholesterol esters, surrounded by a monolayer of amphipathic lipids (phospholipids and free cholesterol), associated with apolipoproteins [8]. Besides functioning as structural components of lipoproteins, apolipoproteins are involved in the metabolism of the lipids they transport by acting as ligands for receptors in cellular membrane, or serving as enzyme regulators. Different lipoproteins contain different classes of apolipoproteins [8, 9]. A schematic figure of lipoprotein structure is presented in figure 2.1.

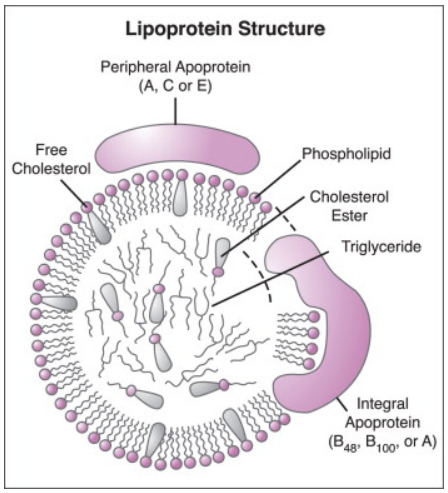


Figure 2.1: Lipoprotein structure (adapted from [9]).

Plasmatic lipoproteins are classified according to its density, which also reflects molecular size by reverse order. By descending order, there are high density lipoproteins (HDL), lipoprotein(a) (Lp(a)), low density lipoproteins (LDL), intermediate density lipoproteins (IDL), very low density lipoproteins (VLDL) and chylomicra [9]. While LDL and HDL have higher content in cholesterol, chylomicra and VLDL present higher content in triglycerides. Total cholesterol (TC) corresponds to the sum of all fractions of lipoprotein's cholesterol [8, 9].

2.1.2 Lipoprotein metabolism

Apolipoprotein B (ApoB) containing lipoproteins comprise the lipid delivery pathway. This is divided in two separate cycles, the exogenous and endogenous lipoprotein pathway [8]. In both cases lipoproteins travel through the organism, becoming smaller and denser, cholesterol enriched particles, as triglycerides are delivered in the form of fatty acids to peripheral tissues.

In the exogenous lipoprotein pathway, dietary lipids and cholesterol are transported through the digestive tract until they reach the intestine, where they are combined with apolipoprotein B-48 (ApoB-48) and A (ApoA) to form immature chylomicra [10,11]. These are then transported through lymphatic vessels into the blood stream, where they will receive apolipoproteins C (ApoC) and E (ApoE) from HDL, originating mature chylomicra [8, 10]. The triglycerides carried by chylomicra will then be hydrolyzed through the action of lipoprotein lipase (LPL) secreted by muscle cells and adipocytes, resulting in the formation of free fatty acids, which can be taken up for either energy production or storage. Along with triglyceride depletion, ApoA and ApoC will also be transferred to HDL molecules in these tissues. As a result of this process, chylomicra remnants will be formed, considerably smaller lipoproteins containing ApoB-48 and ApoE, and enriched in cholesterol esters. Through ApoE, chylomicra remnants will finally bind with hepatic receptors, such as LDL receptor (LDLr) and LDL receptor related protein (LRP), and recycled by the liver [8, 11].

In the endogenous lipoprotein pathway, VLDL are formed in the liver, and will be metabolized in a similar process as chylomicra in the intestine. These particles contain ApoB-100, a different isoform of ApoB, and similarly to chylomicra, will receive ApoE and ApoC from HDL, and will

be transported to peripheral tissues where triglycerides will be hydrolyzed by LPL, transferring ApoC back to HDL at this point. The resulting molecules are named IDL, or VLDL remnants, and contain ApoE and ApoB-100, through which will bind to hepatic receptors. This pathway comprises an additional step however, in which part of IDL molecules suffer further remodeling by hepatic lipase (HepL), transfer ApoE to HDL, and are turned into LDL [8, 11].

LDL has only one apolipoprotein, ApoB-100, and contains most of plasmatic cholesterol. LDL particles are captured in different tissues, through binding of ApoB-100 with LDLr, with special relevance to hepatocytes, similarly to chylomicra and IDL, where they will be hydrolyzed. LDL can also be captured by peripheral cells for its cholesterol content or, of clinical concern, may be taken up by macrophages and some endothelial cells. Through this mechanism, macrophages can accumulate cholesterol on the inside of arterial walls, being referred to as "foam cells". This constitutes the basis of the atherosclerotic process in the early stages [8, 9]. The levels of plasma LDL are determined by the rate of LDL production and clearance, both of which are regulated by the number of LDLr in the liver [10].

Lp(a) is a lipoprotein similar in structure to LDL, differing in the fact that has a unique apolipoprotein, apo(a), attached to Apo-B100. Despite the structural similarities however, Lp(a) synthesis and metabolism is totally independent from LDL. Additionally, evidence suggests Lp(a) is not a metabolic product of other lipoproteins, nor is it metabolized to other lipoproteins. Elevated plasma Lp(a) levels are associated with an increased cardiovascular disease (CVD) risk. Mechanisms by which Lp(a) increases CVD risk are still poorly understood, but include a pro-thrombotic mechanism due to the similarity of apo(a) to the fibrinolytic proenzyme plasminogen, associated with an atherogenic effect mediated by preferential binding to oxidized phospholipids, with enhanced deposition in the artery wall [8].

Apolipoprotein AI (ApoAI) containing lipoproteins, or HDL, participate in reverse cholesterol transport, a process through which excess cholesterol is removed from peripheral cells. HDL originate in the liver and intestine, and in the immature stage have the form of small disks of double lipidic layer, containing phospholipids, cholesterol, and ApoA, C and E [8, 9]. Nascent HDL captures cholesterol from extra hepatic tissues, including macrophages, forming spherical-shaped mature HDL. Mature HDL will then poor its cholesterol content into the liver, either directly, by interaction with hepatic scavenger receptor B1 (SR-B1), or indirectly by transferring the cholesterol to VLDL or LDL [8, 10]. Cholesterol efflux from macrophages to HDL plays an important role in protecting from the development of atherosclerosis, reason why high HDL concentrations are considered to be beneficial [8]. A representation of lipoprotein metabolic pathways is presented in figure 2.2.

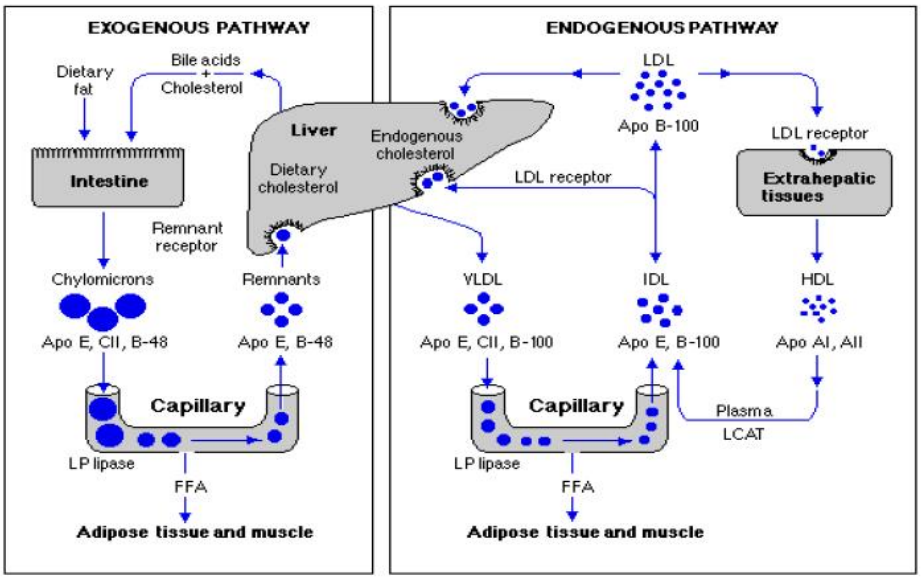


Figure 2.2: Lipoprotein exogenous and endogenous metabolic pathways (adapted from [11]).

2.2 Familial Hypercholesterolaemia

2.2.1 Introduction

FH is a monogenic, autosomal dominant pathology, characterized by elevated plasmatic cholesterol concentrations, in particular LDL cholesterol (LDLc) [1]. These high cholesterol levels from birth lead to its accumulation in arterial walls, promoting the early development of atherosclerosis, and increasing the probability of CVD. Cholesterol accumulation also occurs in extravascular tissues, namely in the cutaneous area, forming tendinous xanthomas which constitute a classical sign of FH [1,3].

This disorder can be divided into its milder heterozygous form (HeFH) and more severe homozygous form (HoFH). While untreated HeFH typically begins to manifest its clinical consequences between the fourth and fifth decade of live, patients with HoFH may suffer significant CV events as early as the first decade of live, and generally do not survive past 30 years of age without therapeutic intervention [1]. Prevalence for HeFH is generally pointed as 1:500, while HoFH is relatively rare, with an estimated prevalence of 1:1,000,000. These values may be highly underestimated however, as prevalence rates as high as 1:200-250 for HeFH [1,7] and 1:160,000-1:300,000 for HoFH have been reported [1]. Since it is the focus of the present study, unless stated otherwise, HeFH will be referred to simply as FH.

2.2.2 Pathophysiology

FH is caused by mutations in three identified genes that encode key proteins involved in the LDLr endocytic and recycling pathways. The etiology of this disease is related to loss-of-function mutations in the LDL receptor gene (*LDLR*) or apolipoprotein B gene (*APOB*), or gain-of-function mutations in proprotein convertase subtilisin kexin type 9 gene (*PCSK9*) [1–3,7]. *LDLR* mutations

are the most frequent, found in >90% of FH subjects, while *APOB* and *PCSK9* mutations are respectively responsible for approximately 5% and 1% of FH cases. The most aggressive phenotype is generally found in *PCSK9* mutations, while less severe FH is usually found with variants in the *APOB* gene [2].

There are over 2300 identified variants in the *LDLR* gene, and over 70% of these are reported to be disease causing, affecting all functional domains of LDLr. A single mutation of the *APOB* gene (p.Arg3527Gln), on the other hand, accounts for most of FH-related cases. Missense *APOB* mutations result in ligand-defective ApoB protein, thus the resulting LDL molecules present reduced affinity to bind to hepatic LDLr. As for the *PCSK9* gene, over 20 different variants have been detected. This gene encodes an enzyme that binds to LDLr resulting in co-internalization and degradation of the receptor within the lysosome. Gain of function mutations in *PCSK9* therefore result in LDL elevation by increased degradation of LDLr, whereas loss of function mutations lead to life-long low LDLc levels and are associated with decreased risk of CVD [12]. As a consequence of the decrease in LDLr functionality and/or availability that results from these genetic mutations, there is a reduction in cellular uptake of LDL and increased plasma LDLc concentration. The consequent cholesterol retention in the arterial wall and foam cell formation within the intima of arteries typically progresses to occlusive atherosclerosis with angina pectoris and/or plaque rupture with coronary heart disease (CHD) [3].

Patients with a clinical diagnosis of FH, with no detectable mutation in one of these three genes may present a polygenic cause for FH, i.e. raised LDLc and TC due to having inherited a greater than average number of common cholesterol-raising variants of modest effect. Alternatively, a monogenic mutation in a novel gene may be present. Several studies have reported that specific mutations in genes like APOE, STAP1, LIPA or PNPLA5 may be causative of a FH phenotype. Prevalence of mutations in these genes is however yet to be determined, and in some cases, pathogenicity to be confirmed [2].

2.2.3 Diagnosis

Early diagnosis of FH has been associated with a significant reduction in CVD risk, supporting the introduction of precocious and/or more aggressive therapeutic measures. Besides lowering CVD--related morbidity and mortality rates, FH early detection and management has also proven to be cost-effective [4]. There are different clinical criteria for FH diagnosis, like Simon Broome (SB) criteria, the Dutch Familial Hypercholesterolemia Diagnostic System or MEDPED criteria [5, 6], although in all cases only genetic testing can positively confirm the diagnostic. SB criteria are the ones adopted in the current study, and take into account family history, presence of physical signs, and TC and LDLc levels. A detailed summary of SB criteria can be found in table 2.1.

One of the major problems presented by the clinical diagnostic is the high false positive rate it presents. This elevated percentage of false positive cases may be partially due to polygenic mutations or monogenic causal mutations not yet identified, as previously referred. A great part of these cases however may be associated with inaccuracy of the clinical diagnosis method [1, 7]. This problem constitutes a heavy burden in terms of healthcare costs, and therefore, the development of new diagnostic methods to increase the rate of true FH cases would be a valuable instrument, allowing

extending the molecular study to a larger universe of subjects, at an earlier stage.

Once an index case is identified based on clinical criteria, and diagnosis is confirmed by the molecular study, a cascade screening process through genetic testing is recommended to identify other family members with FH, which in turn can be referred for adequate therapeutic follow-up [7].

Table 2.1: Simon Broome diagnostic criteria for Familial Hypercholesterolemia [6].

Point	**Criteria**
a) Biochemical indicators	Adults: TC levels >290 mg/dL (>7,5 mmol/L) or LDLc >190 mg/dL (>4,9 mmol/L) Children: TC levels >260 mg/dL (>6,7 mmol/L) or LDLc >155 mg/dL (>4,0 mmol/L)
b) Physical signs	Tendon xanthomas in the patient, or in a first or second degree relative;
c) Molecular study	DNA-based evidence of functional mutation in *LDLR*, *APOB* or *PCSK9* genes;
d) Family history of CVD	Family history of myocardial infarction before age 50 years in a second degree relative or before age 60 years in a first degree relative;
e) Family history of biochemical indicators	Family history of TC >290 mg/dL (>7,5mmol/L) in an adult first or second-degree relative; Family history of TC >260 mg/dL (>6,7mmol/L) in a child or sibling under 16 years of age;
Diagnosis:	
a) + b) or c)	Definite Familial Hypercholesterolemia
a)+ d) or a) + e)	Possible Familial Hypercholesterolemia

TC: total cholesterol; LDLc: low density lipoprotein cholesterol; CVD: cardiovascular disease;

2.2.4 Treatment

The main goal of FH treatment is to lower LDLc levels, thus reducing the risk of atherosclerotic heart disease. Early and aggressive treatment is beneficial, and can substantially reduce the cumulative CVD risk of FH patients [4]. FH patients should undergo a comprehensive treatment program of lifestyle modification, which includes dietary changes, exercise and behavioral therapy. Dietary changes comprise reduction in saturated fats, transfats and cholesterol, and inclusion of foods known to lower LDLc, such as plant sterols and stanols. Risk factors such as hypertension, diabetes, and smoking should be addressed [4, 7]. Although lifestyle modifications are considered beneficial as part of the preventive strategy, these are however unlikely to lower LDLc levels sufficiently, and direct intervention through medication is invariably required [4].

Cholesterol-lowering drugs should be initiated immediately at diagnosis in adults and strongly considered starting at age 8 to 10 years in childhood [4, 7]. Moderate to high potency statins medication constitutes the basis of many treatment regimens, and is generally used as first line treatment. If started prophylactically in early adulthood, statin use has been shown to lower LDLc

levels up to 50%, and CHD risk by up to 80%. A target reduction of $\geq 50\%$ LDLc reduction from baseline has been recommended. The European Atherosclerosis Society (EAS) has outlined LDLc targets of <135 mg/dL for children, <100 mg/dL for adults and <70 mg/dL for adults with CHD or diabetes [7]. Long-term safety of statins in the pediatric population is still unknown, but the current benefits of therapy outweigh the risk of untreated pediatric subjects [4]. Depending on the patient's baseline LDLc levels and responsiveness to therapy, a combination therapy of different medications can be used to treat the FH patient. Drugs that can be added to statins for LDLc reduction include ezetimibe, bile-acid sequestrants, niacin and fibrates [1, 4, 7]. New therapeutic approaches such as PCSK9 inhibitors are also available treatment options, although the potential benefits of its use in children are still under study [4].

Patients with very high CVD risk, whose LDLc levels remain elevated despite combination therapy, may be candidates for LDL apheresis, an extracorporeal treatment that uses various methods to remove LDLc from the circulation [1,4]. Treatment options in severe cases, that are non-responsive to other treatment modalities, include partial ileal bypass and liver transplantation. In children, liver transplantation is restricted to patients with HoFH, and is limited by risks associated to transplant surgery, need for life-long immunosuppression, and limited number of donor livers [4].

2.2.5 The Portuguese FH Study

At a national level, the Portuguese FH Study has been implemented since 1999, with the purpose of diagnosing and stratifying the risk of FH in the Portuguese population, as well as to obtain a deeper understanding of the clinic and molecular mechanisms underlying this pathology [13]. From 1999 to 2017, over 926 index cases have been studied, and 331 patients with a positive molecular diagnosis have been identified. Of the index cases, 389 were children (< 18 years of age), of which 157 tested positive for FH. Relatives of confirmed positive FH index cases have also been tested, following a cascade screening method. A total of 1024 family members have been evaluated through this method, leading to the additional identification of 473 FH individuals. Of the family members, 220 were children, of which 120 were FH positive. This numbers add up to a total of 804 confirmed FH cases. However, even if considering the conservative estimate for the prevalence of this disease of 1:500 individuals [1], this pathology is still severely under diagnosed in the country, with only around 4% of FH carriers identified.

One of the major obstacles in increasing the number of FH confirmed cases is, as referred above, the high false positive rate presented by clinical diagnosis. In the case of the Portuguese FH Study, which uses SB criteria to classify the patients, only around 42% of the individuals with clinical diagnosis presented a positive molecular diagnostic, and although some of these cases may present polygenic mutations or less frequent monogenic mutations not yet identified, a great part is surely due to inaccuracy of the clinical diagnosis [1]. It was in this context that the current study was designed, with the main purpose of developing new tools that can improve the accuracy of FH clinical diagnosis, based on different biochemical markers. An effective increase in the rate of detection of true FH cases by implementation of a new clinical diagnosis method would allow extending molecular testing to a greater population at an early stage, reducing CVD risk through adequate management.

Chapter 3

Methods

The Methods chapter is divided in three main sections. The first two sections comprise a theoretical framework of the classification methods that will be developed in the present study: LR and DT. The third section is dedicated to the description of the methodological procedures used in this work. This last part is divided in sample characterization, presentation of the methods used to collect the biochemical and molecular variables, and statistical procedures, which concern exploratory data analysis, application of LR and DT classification methods, and comparison between classification models through bootstrap resampling techniques.

3.1 Logistic Regression

Logistic regression (LR) is used to analyse the relationship between a dependent or response variable of categorical nature, and one or several independent or predictor variables. In the case of a binary dependent variable, the outcome assumes one of two values, "1" and "0", which may respectively be thought of as "success", whenever the event of interest is observed, and "failure" otherwise. This categorization system can be used to represent any kind of binary outcome, such as pass/fail, win/loose, alive/dead, healthy/sick or positive/negative [14].

The binary LR model is based on the Bernoulli probability distribution, where Y is a random variable that takes values 1 or 0 with probabilities $P(Y = 1) = \pi$ and $P(Y = 0) = 1 - \pi$ [14, 15]. The respective probability mass function can be expressed as

$$f(y_i; \pi_i) = \pi_i^{yi}(1 - \pi_i)^{1-y_i}, \quad \pi_i \in [0, 1],\ y_i = 0, 1. \tag{3.1}$$

Let $\boldsymbol{x}_i$ be the vector representing the value of the independent variables observed for the *ith* individual. The quantity π_i represents the expected value of the outcome variable Y given the value of the independent variables $\boldsymbol{x}_i$, and is represented as $E(Y|\boldsymbol{x}_i)$. Unlike classic linear regression, where $E(Y|\boldsymbol{x}_i)$ is expressed in a linear equation in $\boldsymbol{x}$, and can take any value as $\boldsymbol{x}$ may range between $-\infty$ and $+\infty$, in LR this quantity ranges between 0 and 1, and represents the probability of Y being a "success". In other words, the LR model estimates the probability of the binary response based on one or more predictor variables, also referred to as covariates [14].

In order to keep the mentioned properties of the linear regression model, a transformation is

applied to π_i through a link function, $g(\pi_i)$ [14, 15] . The most commonly used link function in LR is the *logit* transformation, defined as

$$g(\pi_i) = logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x'\beta}, \tag{3.2}$$

where $\boldsymbol{x'\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$.

LR can therefore be seen as a special case of the generalized linear model (GLM) [14]. From this equation, the probability π_i can be calculated, and represented by the logistic function:

$$\pi_i = E(Y|\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x'\beta}}}{1+e^{\boldsymbol{x'\beta}}}, \tag{3.3}$$

Another important difference between linear and LR models concerns the distribution of the error term (ε) of $E(Y|\boldsymbol{x}_i)$. In the linear regression model, an observation of the outcome measure is numerical, and can be expressed as $Y = E(Y|\boldsymbol{x}_i) + \varepsilon$, with ε following a normal distribution of mean zero and constant variance. In LR on the other hand, the value of the outcome measure can only assume two values, and is therefore the binomial model that describes the distribution of the error term ε. In this case $\varepsilon = 1 - \pi_i$ with probability π_i if $Y = 1$, and $\varepsilon = -\pi_i$ with probability $1 - \pi_i$ if $Y = 0$ [14].

Estimates for the $(p+1)$ terms of the vector of parameters $\boldsymbol{\beta}$ are obtained via maximization of the log-likelihood function, which can be expressed as

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)). \tag{3.4}$$

Differentiating the log-likelihood function with respect to $(p+1)$ coefficients will result in a system of $(p+1)$ equations. Because the resulting system does not have an analytical solution, it must be solved using iterative methods, from which the most common is the Iterative Reweighted Least Squares (IRLS) method [14, 15].

3.1.1 Model and Coefficients Significance

The quality of adjustment in LR can be assessed by the likelihood ratio test, using *deviance* (D) test statistic. Deviance is expressed as $-2\log$ of the ratio between the likelihood of the fitted and saturated model. The fitted model is the one built with the selected number of variables, and the saturated model the one containing as many parameters as data points [14].

In LR with a binary outcome variable, it can be proven that the likelihood of the saturated model is identically equal to 1, following that deviance can be obtained as

$$D = -2\log L(fitted\,model) = -2\sum_{i=1}^{n} [y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]. \tag{3.5}$$

The smaller the deviance, the closer the fitted value is to the saturated model. In the opposite sense, the larger the deviance, the poorer the fit.

Deviance measures can also be used to test for overall significance of the LR model. This test,

represented by the statistic G, is calculated as the difference in deviance between the fitted model, and the model including only the intercept, or constant term (null model), and is represented as

$$G = D(null\ model) - D(fitted\ model) = -2log\left[\frac{L(null\ model)}{L(fitted\ model)}\right] \tag{3.6}$$

In an analogous way, G can be used to test if two nested models differ significantly. Considering two models, one containing $(p+v+1)$ parameters, and other containing $(p+1)$ parameters, G test statistic follows an asymptotic chi-square distribution with v degrees of freedom, v being equal to the difference in the number of parameters between the two models [14, 15].

To test the significance of individual parameters, the Wald test is usually preferred over the likelihood ratio test. The Wald statistic (W) is obtained by dividing the maximum likelihood estimate of the parameter by its standard error (SE), which can be estimated by the respective information matrix. Wald test equation is represented as:

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}. \tag{3.7}$$

Under the null hypothesis $H_0 : \beta_j = 0$, this test statistic follows an asymptotic normal standard distribution [14]. The asymptotic $100 \times (1-\alpha)\%$ confidence intervals (CI) for the $\hat{\beta}_j$ can be obtained by calculating

$$\hat{\beta}_j \pm z_{1-\alpha/2} \times \widehat{SE}(\hat{\beta}_j). \tag{3.8}$$

3.1.2 Model Interpretation: Odds and Odds Ratio

One of the great advantages of the LR model is its intuitive relation with the odds and odds ratio (OR) measures [14, 15].

The odds are defined as the probability of the event of interest to occur divided by the probability of it not to occur $(\pi/(1-\pi))$. As can be observed, the logit function presented in the previous section directly corresponds to the logarithm of the odds for the given event (equation 3.2).

In its turn, the OR expresses the relative odds for the occurrence of the outcome of interest, given different exposure to certain variables or risk factors. It is used as a measure of association between the analysed independent variables and the outcome. The OR varies between 0 and $+\infty$. OR values > 1 reflect higher odds for the outcome of interest to occur due to exposure to the independent variables, and the inverse is verified for values < 1.

As an example, the case of just one independent variable of dichotomous nature is illustrated below, which provides a conceptual foundation for all the other situations:

$$OR = \frac{\dfrac{\pi_1}{1-\pi_1}}{\dfrac{\pi_0}{1-\pi_0}} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1), \tag{3.9}$$

where π_1 is $P(Y=1|x=1)$ and π_0 is $P(Y=1|x=0)$. In this case, the $OR = \exp(\beta_1)$ expresses how much more likely it is for the outcome to be present among individuals with $x=1$ comparing

to those with $x = 0$ [14].

The procedure would be analogous for a continuous independent variable, under the assumption that the logit is linear in relation to the covariate. In this case, $\exp(\beta_1)$ corresponds to the OR associated with a one-unit increase in the independent variable. Since many times a value of "1" is not meaningful for a continuous variable, both point estimates and endpoints of the CI can be multiplied by a defined "c" units difference in the covariate. Concerning multivariate data, i.e., a model with multiple independent variables, one can consider that each estimated coefficient provides an estimate of the log-odds adjusting for the other variables in the model, considering the interaction between variables to be non-significant [14, 15].

A $100 \times (1 - \alpha)\%$ CI is used assess the precision of the estimated OR. Since this parameter ranges from 0 to $+\infty$, its estimator, $\widehat{OR}$, tends to have a distribution highly skewed to the right. Therefore, the CI for OR_j is built by first calculating the endpoints of a CI for the $\log OR = \beta_j$, as in equation 3.8 and exponentiating the result [14].

3.1.3 Model Selection

When faced with several potential predictor variables, different procedures are necessary in order to select the subset of variables that best predicts the outcome. Although some model fit measures can improve with variable number, a LR model with many variables is likely to overfit the data, and is characterized by large estimated standard errors [14, 16]. The goal is therefore to select the most parsimonious model, with the best predictive ability.

Variable selection should include several steps. Univariate significance analysis of each candidate independent variable should initially be performed. The adoption of a conservative p-value of 0.20 to retain variables at this point is recommended [14].

The multivariable model should then be fitted, including all identified significant covariates. The selection of variables to retain in the multivariable model can be performed using different procedures. The simplest approach, known as the "Enter" method, is to include all identified covariates in the model simultaneously, and exclude non-significant ones based on Wald test p-value [14, 17].

Alternatively, the designated purposeful selection methods are more sophisticated, and include forward selection, backward elimination, and bidirectional stepwise selection methods. In the forward selection method, variables are introduced one by one, beginning with the most significant, and stopping when addition of the next variable does not significantly improve the quality of the model. In backward elimination method, all variables are initially introduced, and withdrawn one by one based on Wald test statistic, until overall model quality does not deteriorate. Concerning stepwise methods, in forward selection followed by backward elimination, candidate variables are ordered and sequentially included in the null model. As soon as the two first variables are included, and from there forward, there is a backward elimination process, i.e., a check if the model significance is altered by deletion of variables previously included, and the model with higher significance is kept. Inversely, backward elimination followed by forward selection, starts with all variables in the model, and excludes them sequentially. As soon as the first two variables are excluded, a forward selection process takes place, to check if the model significance is altered by inclusion of variables previously excluded, and the model with higher significance is kept [14, 16, 17]. The conservative

p-value threshold of 0.20 is still recommended in multivariable analysis as a threshold to include or exclude variables. Along with statistical significant variables, clinically significant variables should also be taken into account [14].

Following any of the presented selection methods, the fitted model can finally be compared with the full model, i.e. the model containing all candidate predictor variables, using the *deviance* test (see equation 3.6). If different models, using different variable selection methods arise, the Akaike Information Criterion (AIC) is a commonly used measure to compare models with different parameters (non-nested models). This measure is defined as

$$AIC = -2logL(fitted\,model) + 2(p+1), \tag{3.10}$$

where L is the likelihood of the fitted model and p the number of estimated regression coefficients. As noted before, in binary logistic regression $-2logL$ equals to the deviance of the fitted model. This measure therefore represents the model deviance, penalized by the number of parameters. As a general rule, the smaller the AIC the better, reflecting lower loss of information of the model [14,15].

3.1.4 Model Diagnostics I: Residual Analysis

Residuals reflect the difference between fitted and observed values. Several different types of residuals can be calculated, such as raw residuals, Pearson residuals, or deviance residuals [14, 15]. Pearson residuals are obtained by dividing the raw residuals by their estimated SE. In the Binomial model, this is given by

$$p_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}. \tag{3.11}$$

Deviance residuals are based on the deviance, or likelihood ratio chi-squared statistic, and are given by

$$d_i = sgn(y_i - \hat{\pi}_i)\sqrt{-2[y_i log\hat{\pi}_i + (1-y_i)log(1-\hat{\pi}_i)]}. \tag{3.12}$$

The corresponding standardized residuals can be obtained by dividing Pearson or deviance residuals by $\sqrt{1-h_{ii}}$, where h_{ii} is the leverage of the ith observation [15,18]. Similarly to linear regression, h_{ii} is the ith diagonal element of the $n \times n$ estimated hat matrix H. More clearly, leverage measures how far the vector of independent variables deviates from its mean, i.e., how extreme predictor values are. Therefore, standardized residuals correct for both the non-constant variance and the leverage of the observations. Conventionally, observations whose leverage is more than two times the mean leverage values (defined as $(p+1)/n$) should be flagged [18].

The linear relation between a continuous covariate and the logit can be assessed through the plot of standardized residuals against the observed values of the variable [15].

Outliers and Influential Observations

An outlier is an observation whose dependent variable value is unusual given its values on the predictor variables. Graphically, the response does not follow the general pattern set by the remain-

der of the points, thus not fitting the current model [18]. Such observations are characterized by having large residuals.

An influential point is one whose removal from the dataset would cause a large change in the fit. Generally these observations are either outliers or possess high leverage [15, 18]. The influence depends both on the response and explanatory variables.

A popular measure to assess if an observation if influent is Cook's distance, expressed as

$$D_i = \left(\frac{r_i^2}{p+1}\right)\frac{h_{ii}}{1-h_{ii}}, \tag{3.13}$$

where r_i is the Pearson standardized residual value, h_{ii} is the leverage and p the number of explanatory variables in the model [14, 15, 18]. Cook's distance can be therefore seen as a product between leverage and "outlierness". The magnitude of D_i is assessed by comparison with the $F_{(0.5;p+1,n-p-1)}$ quantile. Since $F_{(0.5;p+1,n-p-1)} \approx 1$, points for which $D_i > 1$ are considered influential. Other thresholds to consider an observation as being influential, as $4/N$, or $4/(N-k-1)$, are also referred in the literature [18]. Another useful diagnostic tool for influential observations analysis is the plot of standardized residuals against hat-values.

3.1.5 Model Diagnostics II: Model Adjustment

The model quality of adjustment can be assessed in two different ways: predictive power and and goodness of fit [14].

The coefficient of determination (R^2) is generally used as a measure of the model predictive power. R^2 translates the percentage of the outcome variable variation that can be explained by predictor variables, i.e. how well the dependent variable can be predicted based on independent variables [14, 15]. Logistic regression models do not use true R^2 measures however, since unlike linear regression these are not based on ordinary least squares (OLS), but on likelihood functions, and therefore do not represent the proportion of explained variance, but rather the improvement in model likelihood over a null model. For this reason, these measures are generally referred as pseudo-R^2.

Different pseudo-R^2 measures have been developed. Some of the most often reported methods include McFadden, Cox-Snell and Nagelkerke pseudo-R^2 [19]. McFadden pseudo-R^2 is defined as

$$R^2_{MF} = 1 - \frac{logL(fitted\,model)}{logL(null\,model)}\ , \tag{3.14}$$

where $L(fitted\,model)$ and $L(null\,model)$ denote the log-likelihoods for the model containing the intercept plus the p covariates, and the model containing only the intercept respectively. In its turn, Cox and Snell pseudo-R^2 is calculated as

$$R^2_{CS} = 1 - \left(\frac{L(null\,model)}{L(fitted\,model)}\right)^{2/n}. \tag{3.15}$$

A problem of this pseudo-R^2 measure is that its upper bound is much lower than 1, not fulfilling one of the basic R^2 requirements, which is to vary between 0 and 1 [14]. Nagelkerke presents an

alternative R^2 measure (R^2_N), which is obtained by dividing R^2_{CS} statistic by $(L(null\,model))^{2/n}$, so that the range of possible values extends to 1 [14, 19].

It should be noted that pseudo-R^2 values usually vary significantly, depending on the method used, and that these are typically low when compared with values in linear regression, since each observation needs to be either 0 or 1, but predicted observations are always in between these extremes.

Goodness of fit (GOF) tests on the other hand refer to the accuracy of the probabilities produced by the model, i.e. how well the model fits the set of observations from which it was built. Typically, this is accomplished by comparing the model's fitted values with the observed values [14].

A commonly used test to assess the model GOF with ungrouped data is the Hosmer-Lemeshow (HL) statistic [14, 15]. This test requires the creation of a number of groups based on predicted probabilities. Comparison between observed (O) and expected (E) frequencies is then performed for each group, using Pearson's chi-squared statistic, both for Y=0 and Y=1, and compared with a chi-square distribution with $(g-2)$ degrees of freedom (g = number of groups). A non-significant p-value indicates that fitted values do not differ significantly from observed values, and therefore model fit is good. The test statistic may be calculated as follows:

$$X^2_{HL} = \sum_{k=0}^{1}\sum_{g=1}^{G}\frac{(O_{kg}-E_{kg})^2}{E_{kg}} = \sum_{g=1}^{G}\frac{(O_{1g}-E_{1g})^2}{n_g\overline{\pi}_g(1-\overline{\pi}_g)}. \tag{3.16}$$

HL test has received several critics however, mainly focusing on the fact that test statistic and p-values obtained through this method depend markedly on the number of groups, and although g=10 is generally defined as a guideline, there is no theoretical support to guide this choice [19].

Alternative GOF measures have been provided for this reason, like the one proposed by Le Cessie et al. (LC) [20], that propose a test based on smoothed residuals. Unlike HL test, which relies on the covariate patterns, LC approach is based on each individual observation, and therefore does not rely in dividing up the sample in a certain number of groups. The corresponding test statistic is a weighted sum of the smoothed standardized residuals (rs), and is denoted by

$$\hat{T}_{LC} = \sum_{i=1}^{n}\frac{\hat{r}^2_{si}}{\widehat{var}(\hat{r}_{si})}. \tag{3.17}$$

For small samples, the test statistic follows a chi-square distribution whose degrees of freedom depend on the estimated mean and variance. For big samples the test statistic is well approximated by a normal distribution.

3.1.6 ROC Curve Analysis

A 2x2 contingency table, known as confusion matrix, is a useful way to summarise the results of a LR model [21]. In this confusion matrix, the observed outcome is cross-classified with the predicted outcome, as shown in table 3.1. Different operating characteristics (OC), such as accuracy (Acc), sensitivity (Se), specificity (Spe), positive predictive value (PPV) and negative predictive value (NPV), can then be derived.

Table 3.1: Confusion matrix for a binary outcome (adapted from Fawcett [22]).

	Disease present	Disease absent	Total
Positive test	True positive (TP)	False positive (FP)	TP+FP
Negative test	False negative (FN)	True negative (TN)	FN+TN
Total	TP+FN	FP+TN	N

Selected Operating characteristics:
Accuracy = $(TP + TN)/N$
Sensitivity = $TP/(TP + FN)$
Specificity = $TN/(FP + TN)$
Positive predictive value = $TP/(TP + FP)$
Negative predictive value = $TN/(FN + TN)$

Empirical notions from the different OC can be taken from the equations presented above. *Acc* can be understood as the proportion of true results (either positive or negative) among all assessments. *Se* refers to the proportion of subjects with the disease that show a positive test result, i.e. the proportion of true positive (TP) cases among all cases of disease. *Spe* is the proportion of subjects without the disease that show a negative test result, i.e. the proportion of true negative (TN) cases among all non-diseased. Finally, PPV represents the proportion of TP among all subjects with a positive test result, and NPV the proportion of TN among all subjects with a negative test result [22].

To obtain the predicted outcome from a LR model, it is necessary to define a cutoff probability value that discriminates successes and failures. The cutoff value of 0.5, sometimes assumed by default in statistical *software* tools is not always the best choice. A more complete measure of discriminatory ability involves the analysis of ROC (*Receiver Operating Characteristic*) curves [14, 21].

The ROC curve is an instrument that allows comparing test predictions with the true state of the disease, for each possible cutoff point of the test [22]. Graphically, the ROC curve is a plot of sensitivity versus 1-specificity values over all possible cutpoints (see figure 3.1). The area underneath the curve (AUC) is considered the global index of the test discriminatory ability. The AUC ranges between 0.5 and 1, with higher values corresponding to better discriminatory ability [23]. In the health field, an AUC between 0.8 and 0.9 is generally considered to represent good discriminatory ability for the diagnosis test, whereas an AUC higher than 0.9 represents an excellent discriminatory ability [21].

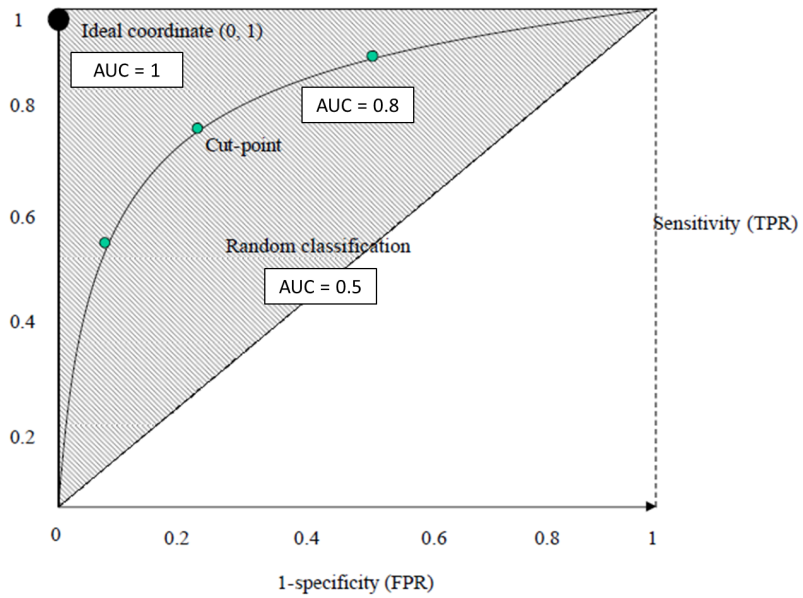


Figure 3.1: ROC curve representation (adapted from Zhu et al. [23]).

Different methods can be employed to select the best possible cut-point, mostly based on ROC curve analysis [24]. A frequently used criterion is to select the point where $Se = Spe$, which corresponds to the point where the product between these two OC is maximum. Graphically, this point is the intersection of the $y = -x$ diagonal line with the ROC curve. Another common approach is to maximize Se and Spe summation, which can be obtained by maximizing Youden's index ($Se + Spe - 1$). This corresponds to the point in the ROC curve with higher vertical distance from the $y = x$ diagonal line (AUC = 0.5). Choosing the nearest point to the point (0,1) in the curve, based on euclidean distance, is another frequent approach. In this case, this will be the point that minimizes $\sqrt{(1-Se)^2 + (1-Spe)^2}$ [24,25]. A different method, based on classical statistics, is the *minimum p-value* approach, which defines the optimal cutpoint as he one that maximizes standard chi-square statistic, calculated for each candidate cutoff value from the respective confusion matrix [25]. Finally, a simple and effective method to determine the optimal cutpoint is through a Two-Graph ROC plot, which represent Se and Spe curves in the same plot, and selects the point where both curves cross as the cutoff value [26].

Although different methods can yield different cutpoints, the purpose is always to achieve a compromise between Se and Spe. The final choice will depend on the purpose of the test, and on which OC are more important in a given situation. Alternatively, Se or Spe threshold values may be predetermined for some reason, like economic costs or emotional consequences associated with diagnosis [21].

3.2 Decision Trees

A decision tree (DT) model consists of a sequence of rules for dividing up a large heterogeneous population into successively smaller, more homogeneous groups, with respect to a particular target

variable. Tree models where the target variable takes a discrete set of values are called classification trees, whereas DT where the target variable is continuous are called regression trees [27]. The present work will focus on the first case, specifically the case of a binary target variable.

Following a tree-like structure, the DT is constituted of nodes, branches and leafs (see figure 3.2). Data enters the tree at the root node, which selects the test that best discriminates among the target classes, according to a discrete function of the predictor variables values, to divide the sample. The branches represent the rule for the split, and each record will be allocated to a respective child node according to this rule. The process is repeated until the records arrive at a leaf node, or terminal node. A leaf node is obtained when no further splitting is possible, or additional splitting does not improve classification. Each leaf node is assigned to a certain class, and although there is a unique path from the root to each leaf, distinct leafs may make the same classification [28]. The process of repeatedly splitting the dataset into smaller subsets, based on an attribute value test, is known as recursive partitioning [27].

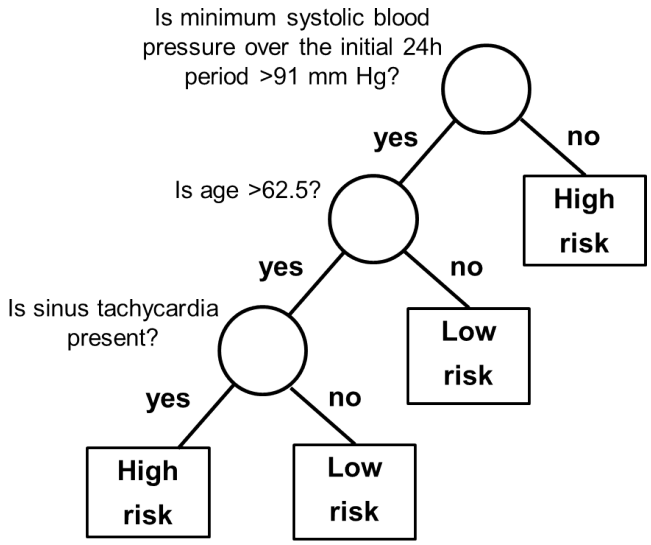


Figure 3.2: Decision tree representation (adapted from Breinman [27]).

Predictor variables in the DT may be categorical or numerical. When splitting on a numerical variable, the midpoint between each two consecutive values for which the outcome of the dependent variable differs is treated as a candidate for the split, and the best cutoff value is then determined [29].

Therefore, splits on a numerical variable take the form $X < z$, where all the the records of the splitting variable X that are less than the cutoff value z are sent to one child node, and all the values greater or equal to z are sent to the other [27, 29].

The DT is built and tested using a set of preclassified data, and is therefore defined as a supervised learning method. After this process, the model can then be used with unclassified data, to assign a record to the most likely class, or to calculate the probability of a given record belonging to a certain class [28].

3.2.1 Entropy Reduction or Information Gain

Like mentioned above, the DT is grown by applying a defined algorithm to a model set comprised of preclassified data. Several algorithms can be used to select the variable and respective cutoff value (in case of a numerical predictor) at each step that best splits the dataset concerning the target variable, i.e. that reveals the most increase in purity of the target variable within the subset. Different metrics such as information gain, Gini impurity index or variance reduction are used [28]. The current work will focus on the first.

Information gain is based on the concept of entropy, which derives from information theory. In general terms, entropy is a measure of how disorganized a system is. It represents the number of bits required to describe a particular outcome, which of course will depend on the size of the set of possible outcomes [27,28]. In other words, it can be thought of as the number of yes/no questions it would take to determine the state of the system. Mathematically, entropy is defined as

$$H(D) = \sum_{c=c_1}^{c_k} (-p_c \cdot \log_2(p_c)), \tag{3.18}$$

where:

D = Set of elements to classify;
$c_1, ..., c_k$ = Classes of the target variable;
p_c = Proportion of elements of c class in D, $c = c_1, ..., c_k$.

When entropy reduction is chosen as a splitting criterion, the algorithm searches for the split that reduces entropy by the greatest amount, i.e., where the information gain is higher. Therefore, information gain and entropy reduction are interchangeable terms. Information gain is used to decide which feature to split on at each step in building the tree, in order to increase the purity of the child nodes by the greatest amount possible [27,29]. The respective equation is represented as follows:

$$Gain(D, A) = H(D) - \sum_{a=a_1}^{a_m} \left(\frac{|D_a|}{|D|} \cdot H(D_a) \right), \tag{3.19}$$

where:

D = Set of elements to classify;
$|D|$ = number of elements in D;
A = Attribute, or predictor variable;
$a_1, ..., a_m$ = Splitting criteria, based on a value of the attribute A;
Da = Subset of D resulting of the application of splitting criteria a, $a = a_1, ..., a_m$;
$|Da|$ = number of elements in Da.

Finally, this algorithm can be implemented in iterative fashion so that, at each split, it searches for the variable and respective cutoff value with higher information gain, until the tree is fully grown, or criteria to stop the algorithm are met [29].

3.2.2 Accuracy Estimation

Unless the target variable is completely separable within a certain number of partitions based on predictor variables, the DT cannot provide a 100% accurate classification rate. Let $R^*(d)$ be the true misclassification rate, or true error rate of a classifier $d(x)$, a real-valued function of the vector of predictor variables $\boldsymbol{X}$. In other words, $R^*(d)$ can be understood as the probability that the classification tree will misclassify a record providing from the same distribution as the one used to build the tree [27].

Different methods can be used to estimate $R^*(d)$. When the dataset is big enough, the original dataset can be divided in two different samples, the training set and the testing set. While the training set is used to grow the tree, $d(x)$, the testing set is posteriorly run through the model to estimate $R^*(d)$. This method is known as *test sample estimation*, and a general rule to obtain the training and testing sets is to randomly divide the original sample in proportions of 2/3 and 1/3, respectively [27,28].

A major problem in most studies however is the fact that datasets are generally of limited dimension, and additional independent records are difficult to obtain. In such cases, data in the learning sample (L), must be used both to construct $d(x)$ and estimate $R^*(d)$. Different types of estimates can be made in this case. The most common and inaccurate method is to directly obtain the misclassification rate from the sample used to grow the tree. This rate, known as *resubstitution estimate*, and represented as $R(d)$, is a very poor estimate of $R^*(d)$, as it leads to severe overfitting. For this reason, resampling methods, like *bootstrap resampling* or v-fold cross validation are preferred when working with smaller sample sizes. Both these methods use resampling from L in order to obtain additional information about the fitted model. In the first case, m samples are generated by random sampling with replacement from L. The different samples are then run through the model, and the average error of classification is used to estimate $R^*(d)$ [30]. An explanation of bootstrap resampling methods with greater detail is presented in the next section. In the second case, L is divided into v subsets of nearly equal size, $L_1, ..., L_v$, and for every v, the model algorithm is then applied using $L - L_v$ as training sample, and L_v as testing sample. The cross validation misclassification rate is then obtained by averaging the misclassification rates of all v models, and the final classifier is constructed using L. [27].

3.2.3 Obtaining the Right Size Tree: Pruning

As mentioned before, the DT can be grown until no additional splits can be made to the training set, or additional splitting does not result in further information gain. In an extreme case, the tree may arise to entirely pure leaf nodes, each one containing just a single element, with a corresponding resubstitution estimate, $R(d)$, equal to zero. The full grown tree however, is generally not the one that performs better classifying a new set of records, since it overfits the training set, thus enhancing the true misclassification rate $R^*(d)$. Very small trees on the other hand will also present an inflated $R^*(d)$, since they will not use some of the available information in L. The extent to which splits are truly informative is therefore questionable, and presents a crucial challenge in tree structured procedures. This problem is generally presented in literature as the bias-variance trade-off [27,28]. The same problem occurs in the LR model presented before. In that case, variables are introduced

or removed sequentially, and the fit is stopped when the deviance test fails to achieve a predefined α level (see subsection 3.1.3).

A procedure created to deal with this problem, called pruning, consists in growing the full or nearly full tree, and then upwards eliminate the smaller nodes that provide the least predictive power, selecting a subtree that will perform much better with new data records. There are several pruning algorithms available. In the present work, bootstrap resampling methods will be used to compare the DT of different size, and select the one with lowest estimated misclassification rate, or higher accuracy.

3.3 Bootstrap Resampling

In many statistical problems, it is important to have not only a point estimate of a certain parameter of interest θ, but also an idea of the variability associated with the estimate. A possible way to estimate this variability is to draw repeated samples from the population of interest, and observe how the statistic of interest fluctuates among the several samples. All the possible values of the sample statistic can then be presented in the form of a probability distribution, called a sampling distribution [31]. To draw several different samples from a certain population is however an expensive and time consuming process, and alternative methods have been developed to solve this problem.

Bootstrap resampling, also referred to as bootstrapping, is a statistical method for estimating the sampling distribution of an estimator T, by sampling with replacement from the original sample [30,31]. Simply described, sampling with replacement means that after an observation is randomly drawn from the original sample, it is put back into the sample before drawing the next observation. As a result, each observation can be drawn once, more than once, or not at all [30]. Each bootstrap sample assumes the same size N as the original sample, and m bootstrap samples can be drawn following this process. The statistic of interest can then be calculated for each bootstrap sample, and the bootstrap distribution used as an estimate of the sampling distribution [31]. The basic idea behind the bootstrap resampling method is therefore to treat the study sample as the population of interest, and repeatedly resample from this specific set of data to obtain an estimate of the sampling distribution.

Bootstrapping methods can be divided into parametric or non parametric. While parametric bootstrapping assumes the sample data $\mathbf{x} = (x_1, x_2, ..., x_n)$, comes from a fully specified distribution model, non parametric bootstrapping makes no other assumption than the fact the random variables X_j are independent and identically distributed [31]. Non parametric methods were the ones used in the present study. Bootstrap resampling techniques can be used for a variety of purposes, such as estimating the error associated with a certain estimate of θ, the respective $100 \times (1-\alpha)\%$ CI, or to conduct hypothesis testing, among others [31,32].

3.4 Methodological Procedures in the Study

3.4.1 Sample

The sample used in this study was constituted by 252 pediatric patients, participants in the Portuguese FH Study, previously presented. Index cases of both sexes, with ages between 2 and 17 years, meeting the clinical criteria for dyslipidemia [33], and not under hypolipidemic medication during the evaluation period were included. Cases presenting an unknown mutation, polygenic mutation, or homozygous FH were excluded, thus limiting the scope of this work to HeFH. All participants had an informed consent form signed by the legal guardian (see Appendix A), and information was registered in a confidential database, legalized by the National Data Protection Commission.

3.4.2 Blood samples collection and processing

Blood samples for DNA extraction and biochemical panel determination were collected in an EDTA and gel tube (10 mL and 7.5 mL respectively). Previous to blood samples collection, during the initial consultation, study participants were submitted to a physical exam and information concerning family history, comorbidities, medication, social-economic and lifestyle indicators was registered (see Appendix B).

For the determination of biochemical indicators, blood serum concentration of different lipidic parameters were used, specifically TC, LDLc, HDLc, TG, Lp(a), ApoAI and ApoB. These were determined by enzymatic and colorimetric methods, using a Cobas Integra 400 Plus (Roche) analyzer. Concentrations were determined in mg/dL, and obtained by technicians from the laboratory of Clinical Chemistry from the Integrated Laboratory Unit at Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA). All samples were received by mail or collected at INSA, and stored at -80ºC until posterior use.

Genomic DNA was extracted from leucocytes of a 5 mL sample of peripheral blood, using a Wizard(r) Genomic DNA Purification Kit (Promega) according to manufacturer's instructions. The molecular study for FH was conducted in three stages. In stage I, mutations were searched in the 18 exons and promotor of *LDLR* gene, and exons 26 and 29 of *APOB* gene, binding sites to LDLr when translated. This step was performed through fragment amplification by polymerase chain reaction (PCR), followed by direct sequencing using Sanger's method. In stage II, gene rearrangements in *LDLR* gene were searched through multiplex ligation-dependent probe amplification (MLPA) technique, in the index cases with no found mutation in the previous step, or with identified mutation but aggressive phenotype. Step III consisted in the molecular study of 12 exons in *PCSK9* gene where FH causing mutations are described, in the index cases with no found mutation in steps I and II, through polymerase chain reaction (PCR) amplification and direct sequencing using Sanger's method [34]. In participants where molecular diagnostic revealed an unknown alteration in one of these genes, functional studies were conducted to verify its pathogenicity. In case no molecular alteration was observed, the participant was classified as negative for FH.

3.4.3 Statistical Procedures

Exploratory analysis of personal characteristics and biochemical profile of individuals with and without FH was initially performed. Respective density plots were drawn, and continuous variables between the FH and non-FH groups were tested by non-parametric methods, using Kolmogorov-Smirnov (KS) test, whereas categorical data was assessed by the chi-square test, adopting a significance level $\alpha = 0.01$.

Classification rules were built using two different approaches: A LR model, based on classical inference methods, and a DT model, a data mining approach based on information theory. Purposeful selection methods were employed to fit the LR model. Residual analysis was performed and the model was fit with and without potential influential observations. Selection of the best cutoff point to use LR as a classification model was performed through ROC curve analysis, adopting two different methods: Youden's index, and *minimum p-value* approach. A confusion matrix was obtained for each of the cutoff points, and different OC were estimated: Acc, Se, Spe, PPV and NPV [22,35]. Different pseudo R^2 and GOF measures were also assessed. Selection between the LR models with and without potential influential observations was performed by bootstrap resampling methods.

For the DT model, entropy and information gain measures were calculated for each biochemical marker, and variables were ranked accordingly [29]. A modified version of the DT model was implemented, based in the sequential exclusion of predictor variables as they are used in each tree node. Following this procedure, the variable with highest information gain was used in the initial node, and new entropy measures were calculated for the remaining variables in the following node, repeating this procedure throughout the tree. All this process had to be manually implemented in R, through the development of the respective functions and cycles, since the pre-existing packages, which provide already developed DT algorithms, do not contemplate this possibility of variable sequential exclusion. An extract of the R code developed for this purpose is presented in Appendix C. The motivation behind this approach was to produce a classification rule that would resemble typical medical criteria, which usually consider single cutpoints. The final tree was pruned to avoid overfitting, by comparing the classification error of DT with different node number, through bootstrap resampling techniques. A confusion matrix and respective OC were also obtained for the final DT.

The LR and DT models were finally compared with the biochemical markers used in SB criteria for FH diagnosis in pediatric subjects, also by means of bootstrap resampling methods. A total of 200 bootstrap samples were generated from the original dataset. This is considered to be an appropriate number, since the true misclassification rate estimate is demonstrated to stabilize from 100 bootstrap samples [36]. These samples were the same used when selecting the respective LR and DT model, in order to allow a global comparison of results. For each bootstrap sample, a confusion matrix was generated using each classification method, by comparison with molecular study results, and correspondent median and mean values of OC were used for performance comparison. Significant differences in the performance of each model, concerning these OC, were assessed by Wilcoxon signed rank test. Concordance between the final models regarding the original sample was also investigated. Statistical analysis was performed using R and R Studio software.

Chapter 4

Results

The results obtained in this study are presented in the current chapter. Exploratory data analysis for sample characteristics and biochemical indicators is initially performed, with significant differences in these variables between FH and non-FH groups being reported. The different classification methods are then developed, using sample data as the training set: SB criteria, the LR model and the DT model respectively. Bootstrap resampling techniques are used to select the best LR and DT models, as well as to compare the predictive performance between the different classification methods, using mean and median values of several OC. Concordance between the final models in classifying the original sample, by comparison with molecular study results, is finally reported.

4.1 Exploratory analysis

A summary description of sample characteristics is presented in table 4.1. No significant differences were found regarding the variables gender or age, between FH and non-FH individuals.

Table 4.1: Sample characteristics regarding number of participants, gender, age at diagnosis and affected gene in FH cases.

	FH	non-FH	p-value
N(%)	83 (32.9)	169 (67.1)	
Male; n(%)	39 (47.0)	61 (36.1)	0.097
Age; mean (sd)	9.31 (3.96)	9.75 (3.57)	
Age group; n(%):			
2-7 years	27 (10.7)	41 (16.3)	
8-12 years	40 (15.9)	90 (35.7)	0.376
13-17 years	16 (6.3)	38 (15.1)	
Gene; n(%):			
LDLR	78 (94.0)		
APOB	5 (6.0)		

* p-value for gender and age group differences calculated using chi-square test. FH: familial hypercholesterolemia

To explore the differences in biochemical variables distribution between FH and non-FH individuals, respective density plots have been obtained (figure 4.1).

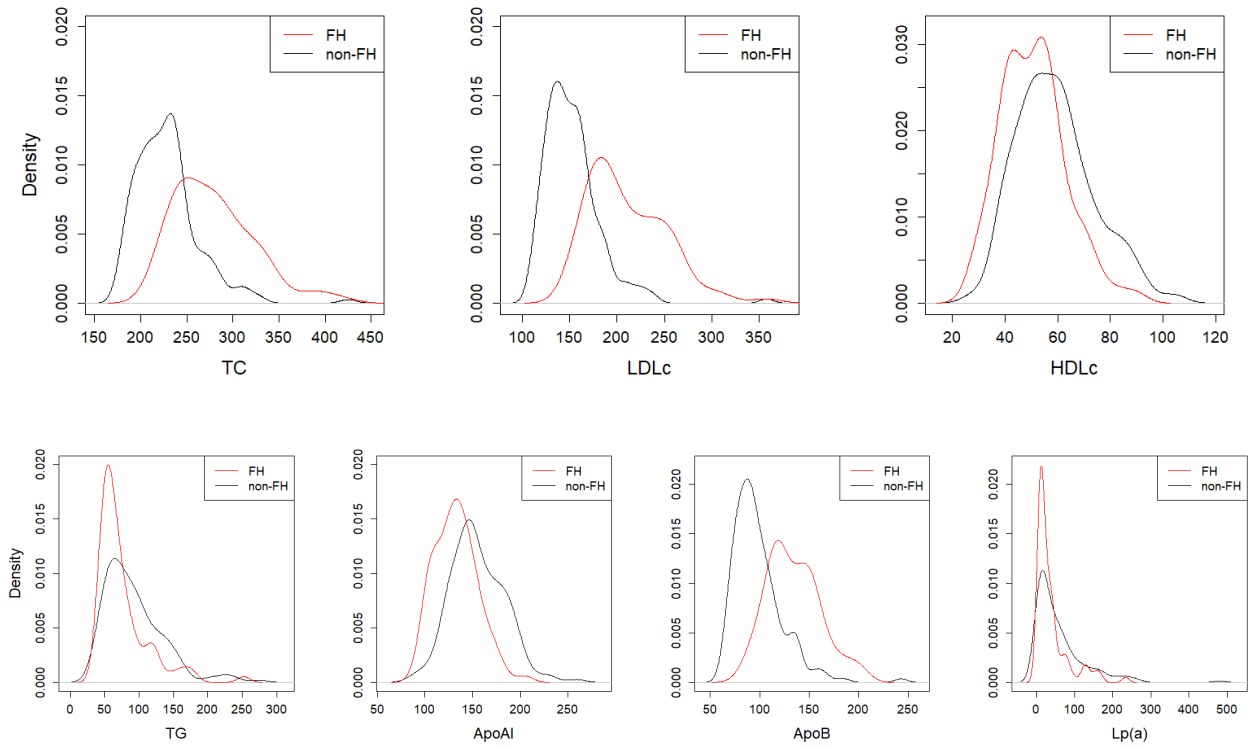


Figure 4.1: Density plots for the biochemical variables between FH and non-FH patients. FH: familial hypercholesterolemia; TC: total cholesterol; LDLc: low density lipoprotein cholesterol; HDLc: high density lipoprotein cholesterol; TG: triglycerides; Apo: apolipoptrotein; Lp(a): lipoprotein(a).

From the observation of the graphics presented above, it is possible to detect apparent differences in the respective density curves for several biochemical variables, suggesting presence of significant differences between FH and non-FH individuals. The graphs also suggest a lack of adjustment to normal distribution. The one sample Kolmogorov-Smirnov (KS) test with Liliefors correction for normality assessment corroborates these findings, with the hypothesis of normality being rejected for all variables in the non-FH group, and LDLc, TG and Lp(a) in the FH group ($p < 0.01$). Non-significant values were only verified among the FH group, for TC ($p = 0.099$), HDLc ($p = 0.237$), ApoAI ($p = 0.582$) and ApoB ($p = 0.121$) variables. For this reason, the non-parametric KS test for two independent samples was chosen to check for significant differences between FH and non-FH groups, concerning these variables. Summary descriptive statistics, along with respective p-values for the mentioned tests, are presented in table 4.2.

Table 4.2: Plasmatic concentrations for biochemical variables in FH and non-FH participants.

	FH					non-FH					p-value
	median	mean	sd	min	max	median	mean	sd	min	max	
TC	274.00	279.51	43.93	214.0	419.0	225.00	228.60	34.33	179.0	425.0	<0.001
LDLc	197.00	210.70	41.35	149.0	358.0	147.00	152.59	30.29	111.0	358.0	<0.001
HDLc	51.00	50.83	12.41	27.0	90.0	58.00	59.15	14.85	27.0	106.0	0.001
TG	65.00	75.90	37.34	39.0	252.0	84.00	92.55	43.33	34.0	275.0	0.001
apoAI	132.00	132.73	22.40	90.0	205.0	150.00	155.34	28.03	86.0	259.0	<0.001
apoB	131.00	134.56	26.99	82.0	205.0	93.00	98.49	25.24	63.0	243.0	<0.001
Lp(a)	20.80	38.69	44.20	1.7	234.0	35.00	59.79	67.22	8.3	480.0	0.006

* All concentrations are expressed in mg/dL; ** p-values calculated using two-sample Kolmogorov-Smirnov test. FH: familial hypercholesterolemia; TC: total cholesterol; LDLc: low density lipoprotein cholesterol; HDLc: high density lipoprotein cholesterol; TG: triglycerides; Apo: apolipoptrotein; Lp(a): lipoprotein(a).

4.2 Simon Broome Criteria

The application of SB biochemical criteria for FH diagnosis to the study sample can be presented in the form of a one node DT, like the one produced in figure 4.2. Respective confusion matrix and operating characteristics are presented in table 4.3.

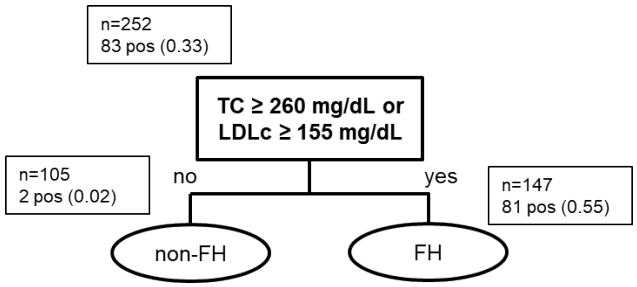


Figure 4.2: Simon Broome biochemical criteria application in the study sample. Number of participants and familial hypercholesterolemia (FH) cases are presented in text boxes. TC: total cholesterol; LDLc: low density lipoprotein cholesterol.

Table 4.3: Confusion matrix and respective operating characteristics for the Simon Broome criteria applied to the study sample.

		Molecular test result		
		Positive	Negative	Total
SB criteria	Positive test	81	66	147
	Negative test	2	103	105
	Total	83	169	252

Op. char.	
Acc	0.73
Se	0.98
Spe	0.61
PPV	0.55
NPV	0.98

SB: Simon Broome; *Acc*: accuracy; *Se*: sensitivity; *Spe*: specificity; *PPV*: positive predictive value; *NPV*: negative predictive value.

4.3 Logistic Regression Model

The development of the LR model followed several steps. Previous to model development itself, a variance inflation factor (VIF) analysis was performed. In this analysis, variables with VIF > 4, which expresses elevated multicollinearity, were sequentially eliminated. The elimination process starts with the variable with higher VIF, and ends when all variables achieve a VIF < 4 (see table 4.4).

Table 4.4: Sequential variable elimination for VIF > 4, in the complete sample ($N = 252$).

	Age	Gender	TC	LDLc	HDLc	TG	ApoAI	ApoB	Lpa
9 vars	1.14	1.04	25.30	23.04	6.73	2.16	3.04	2.70	1.07
8 vars	1.12	1.04	-	2.72	3.22	1.43	3.01	2.72	1.05

TC: total cholesterol; LDLc: low density lipoprotein cholesterol; HDLc: high density lipoprotein cholesterol; TG: triglycerides; Apo: apolipoptrotein; Lp(a): lipoprotein(a).

As can be seen in the table above, all variables presented VIF < 4 after the removal of the first variable, TC, and the model with 8 variables was therefore considered as the full model. In spite the fact the variables age and gender did not present a significant relationship with the presence of FH, as presented in the exploratory analysis conducted previously, these were still included in the LR model so that the outcome could be controlled for eventual confounding by these factors.

Variable selection for the final model was performed using different purposeful selection procedures: forward, backward, stepwise forward and stepwise backward methods. In all cases, the obtained final model contained the same variables and Akaike information criterion (AIC) value, equal to 176.78. The model characteristics are presented in table 4.5:

Table 4.5: Final model fit for the biochemical variables, in the complete sample ($N = 252$).

	β_j	SE	Wald	p-value	OR	95% CI
(Intercept)	-3.665	1.704	-2.150	0.032	0.03	(0.00 - 0.69)
LDLc	0.053	0.008	6.825	<0.001	1.05	(1.04 - 1.07)
TG	-0.023	0.006	-3.847	<0.001	0.98	(0.97 - 0.99)
ApoAI	-0.029	0.008	-3.606	<0.001	0.97	(0.96 - 0.99)
Lpa	-0.008	0.004	-1.890	0.059	0.99	(0.98 - 1.00)

SE: standard error; OR: odds ratio; CI: confidence interval; LDLc: low density lipoproteins; TG: triglycerides; ApoAI: apolipoptroteinAI; Lp(a): lipoprotein(a).

As can be observed, the final model included the variables LDLc, TG, ApoAI and Lp(a). All variables were significant for $\alpha = 0.05$, except Lp(a), which presented a borderline value $p = 0.059$. In a posterior analysis of Lp(a) variable individually, a significant p-value was found, both for the Wald test ($p = 0.012$) and for analysis of Deviance, when comparing the models with and without this variable ($p = 0.047$). For this reason, the variable was left in the model at this point. Accordingly, the logit function was estimated as

$$\widehat{g(\pi)} = -3.665 + 0.053 \times LDLc - 0.023 \times TG - 0.029 \times ApoAI - 0.008 \times Lp(a), \tag{4.1}$$

where π represents the probability of the individual to have FH.

4.3.1 Residual Analysis

Residual analysis was accomplished through graphical methods, namely plots for standardized residuals against linear predictor, outliers identification, leverage, Cook's distance and influential observations plots, which are presented in figure 4.3.

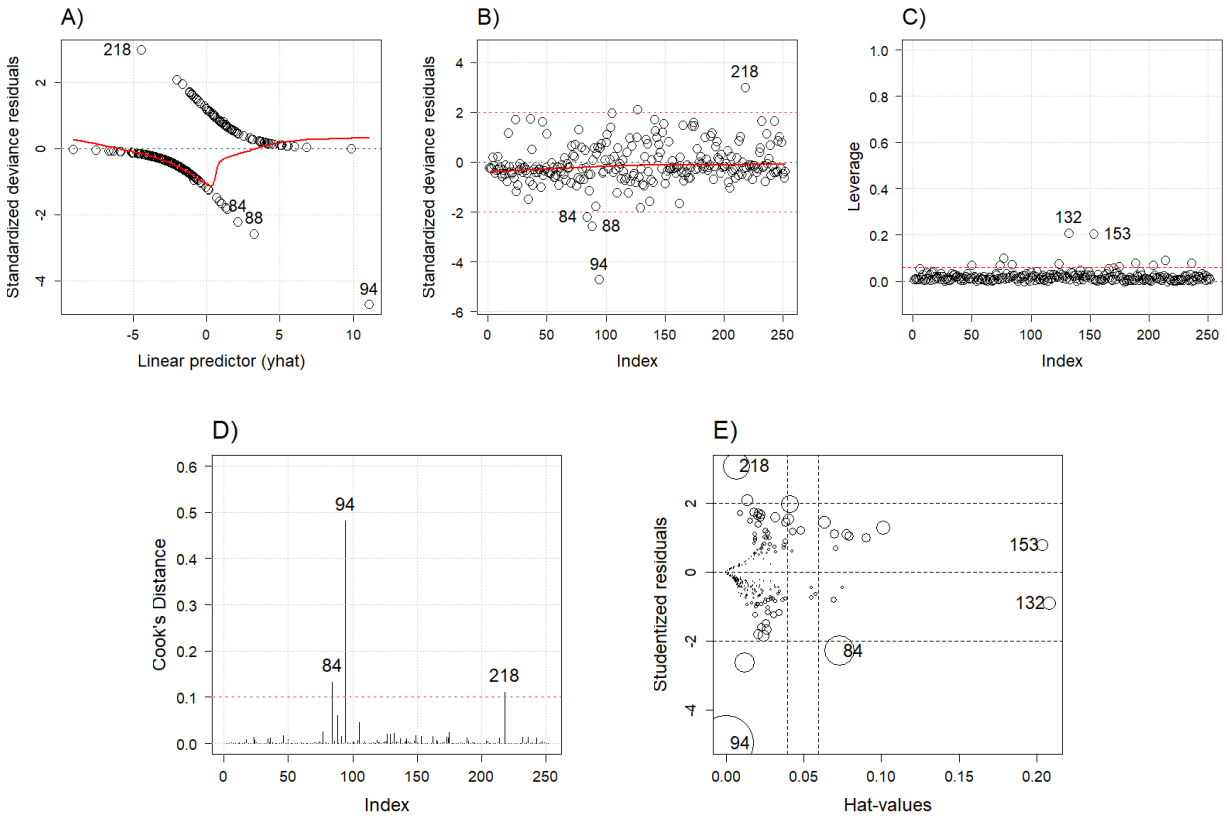


Figure 4.3: Residual analysis plots: A) Standardized residuals vs linear predictor; B) Outliers analysis; C) Leverage analysis; D) Cook's distance plot; E) Influential observations plot.

From these results, a total of 5 observations were signalled as potentially influential (observations number 84, 94, 132, 153 and 218). These observations were then excluded from the dataset, and the entire variable selection process was repeated for this sample ($N = 247$). As before, the process started by VIF examination, as shown in table 4.6.

Table 4.6: Sequential variable elimination for VIFs > 4, of data without influential observations ($N = 247$).

	Age	Gender	TC	LDLc	HDLc	TG	ApoAI	ApoB	Lpa
9 vars	1.12	1.05	18.81	14.67	6.13	2.43	2.79	2.11	1.11
8 vars	1.09	1.05	-	2.22	2.89	1.74	2.79	2.12	1.10

TC: total cholesterol; LDLc: low density lipoprotein cholesterol; HDLc: high density lipoprotein cholesterol; TG: triglycerides; Apo: apolipoptrotein; Lp(a): lipoprotein(a).

Compared to the previous table, generally lower VIF values have been produced, with TC still showing high multicollinearity with other biochemical variables, and therefore removed. Variable selection for the final model was performed following the four purposeful selection methods already mentioned, and final model characteristics can be observed in table 4.7.

Table 4.7: Final model fit for the biochemical variables, of data without influential observations ($N = 247$).

	Estimate	SE	z value	p-value	OR	95 % CI
(Intercept)	-7.083	2.252	-3.146	0.002	0.00	(0.00 - 0.06)
LDLc	0.086	0.013	6.631	<0.001	1.09	(1.07 - 1.12)
TG	-0.041	0.009	-4.536	<0.001	0.96	(0.94 - 0.98)
ApoAI	-0.037	0.010	-3.761	<0.001	0.96	(0.94 - 0.98)

SE: standard error; OR: odds ratio; CI: confidence interval; LDLc: low density lipoprotein cholesterol; TG: triglycerides; ApoAI: apolipoptroteinAI.

As in the model using all observations (equation 4.1), the final model was also the same regardless of the selection procedure, presenting an AIC of 131.90, a lower value than before. In this case however, the final model included only the variables LDLc, TG and ApoAI, all significant for $\alpha = 0.05$. The corresponding logit function was therefore defined as

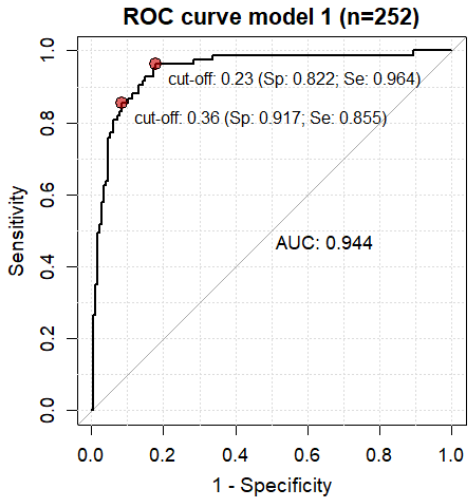
$$\widehat{g(\pi)} = -7.083 + 0.086 \times LDLc - 0.041 \times TG - 0.037 \times ApoAI. \tag{4.2}$$

A more thorough analysis of the variable Lp(a), now excluded, has shown that this biochemical parameter is still significant when considered alone, as assessed by the Wald test ($p = 0.023$). However, the deviance test between this new model, and a model including Lp(a), using the sample without the identified influential observations, reveals this variable is no longer significant now ($p = 0.598$). Further analysis was performed for both models, in order to try to understand the main differences in terms of quality and predictive ability between the two.

4.3.2 Model adjustment

ROC curve analysis was performed for both LR models, and corresponding goodness of fit measures were estimated. From ROC curves, selection of the best cutoff value was performed by two different approaches, Yoden index and minimum p-value methods, and respective OC were calculated for each cutoff point. The entire process is presented in figure 4.4. It is important to remind however that, while the initial LR model (model 1) used all subjects data (including potential influential observations), the second LR model (model 2) excluded these cases, and therefore a direct comparison between their performance cannot be made.

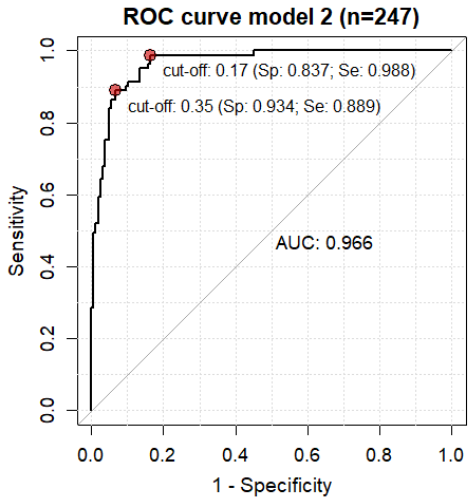
A)



pseudo-R^2 measures	
Cox-Snell R^2	0.465
Nagelkerke R^2	0.647
McFadden R^2	0.494
GOF tests	
HL chi-square = 17.292	p-value = 0.027
LC z-statistic = -5.065	p-value $<$ 0.001

cutoff value	Confusion Matrix				Operating characteristics				
	TP	FP	TN	FN	*Acc*	*Se*	*Spe*	*PPV*	*NPV*
c=0.23	80	30	139	3	0.87	0.96	0.82	0.73	0.98
c=0.36	71	14	155	12	0.90	0.86	0.92	0.84	0.93

B)



pseudo-R^2 measures	
Cox-Snell R^2	0.551
Nagelkerke R^2	0.767
McFadden R^2	0.632
GOF tests	
HL chi-square = 28.448	p-value = $<$ 0.001
LC z-statistic = -0.446	p-value = 0.656

cutoff value	Confusion Matrix				Operating characteristics				
	TP	FP	TN	FN	*Acc*	*Se*	*Spe*	*PPV*	*NPV*
c=0.17	80	27	139	1	0.89	0.99	0.84	0.75	0.99
c=0.35	72	11	155	9	0.92	0.89	0.93	0.87	0.95

Figure 4.4: ROC curve plots with respective cutoff values for models with all observations, and without influential observations. Pseudo-R^2 and GOF measures, and a confusion matrix with operating characteristics calculated for each cutpoint are presented, in the tables at the right and bellow the figure respectively. TP: true positive; FP: false positive; TN: true negative; FN: false negative; *Acc*: accuracy; *Se*: sensitivity; *Spe*: specificity; *PPV*: positive predictive value; *NPV*: negative predictive value; HL: Hosmer-Lemeshow; LC: Le Cessie.

4.3.3 Selection between the two LR models

In order to compare LR model 1 (LR1) with LR model 2 (LR2), bootstrap resampling techniques were applied. Using 200 bootstrap samples of the complete sample ($N = 252$), the performance of both LR models was compared through analysis of corresponding OC. For each model, the two cutoff values previously determined have been used ($c = 0.23$ and $c = 0.36$ for LR1 and $c = 0.17$ and $c = 0.35$ for LR2). Results can be seen in figure 4.5 and table 4.8.

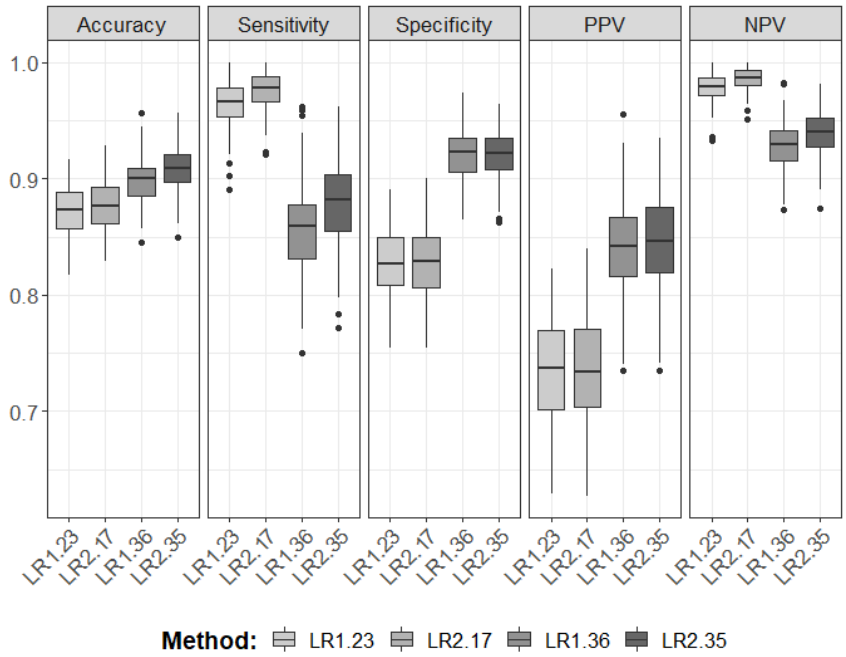


Figure 4.5: Boxplots representing the performance of the selected cutpoints in LR1 and LR2 models in different operating characteristics, over 200 bootstrap samples. LR: logistic regression.

Table 4.8: Descriptive statistics for operating characteristics in the selected cutpoints in LR1 and LR2 models, over 200 bootstrap samples.

	TP	FP	TN	FN	*Acc*	*Se*	*Spe*	*PPV*	*NPV*
				LR1.23					
Median	80.00	29.00	140.00	3.00	0.873	0.966	0.827	0.738	0.980
Mean	80.17	29.12	139.83	2.87	0.873	0.965	0.828	0.734	0.980
sd	7.17	5.00	7.33	1.65	0.021	0.020	0.028	0.042	0.011
min	62	19	122	0	0.818	0.890	0.754	0.629	0.932
max	98	43	159	10	0.917	1.000	0.891	0.822	1.000
$Q_{0.25}$	75.75	25.00	134.00	2.00	0.857	0.954	0.809	0.702	0.972
$Q_{0.75}$	85.00	33.00	145.00	4.00	0.889	0.978	0.850	0.769	0.986
				LR1.36					
Median	71.00	13.00	155.50	12.00	0.901	0.859	0.923	0.842	0.930
Mean	71.05	13.47	155.49	11.99	0.899	0.856	0.920	0.841	0.928
sd	6.90	3.57	7.55	3.33	0.019	0.038	0.021	0.039	0.020
min	54	4	137	3	0.845	0.750	0.864	0.735	0.874
max	89	23	174	21	0.956	0.962	0.974	0.956	0.983
$Q_{0.25}$	66.75	11.00	151.00	10.00	0.885	0.831	0.905	0.816	0.916
$Q_{0.75}$	75.00	16.00	160.00	14.00	0.909	0.878	0.935	0.867	0.941
				LR2.17					
Median	81.00	29.00	140.00	2.00	0.877	0.978	0.829	0.733	0.986
Mean	81.20	29.04	139.91	1.84	0.877	0.978	0.828	0.737	0.987
sd	7.17	5.15	7.83	1.35	0.021	0.016	0.030	0.042	0.009
min	63	17	119	0	0.829	0.921	0.754	0.627	0.951
max	98	42	159	7	0.929	1.000	0.900	0.839	1.000
$Q_{0.25}$	76.00	25.00	134.00	1.00	0.861	0.967	0.806	0.704	0.980
$Q_{0.75}$	86.00	33.00	145.00	3.00	0.893	0.988	0.850	0.770	0.993
				LR2.35					
Median	73.00	13.00	156.00	10.00	0.909	0.882	0.921	0.847	0.940
Mean	73.04	13.49	155.47	10.01	0.907	0.879	0.920	0.844	0.940
sd	6.97	3.63	7.52	2.99	0.019	0.034	0.021	0.039	0.018
min	57	6	136	3	0.849	0.772	0.863	0.735	0.874
max	93	24	176	21	0.956	0.962	0.964	0.935	0.981
$Q_{0.25}$	68.00	11.00	150.00	8.00	0.897	0.855	0.908	0.819	0.927
$Q_{0.75}$	77.00	15.25	160.00	12.00	0.921	0.903	0.934	0.875	0.952

TP: true positive; FP: false positive; TN: true negative; FN: false negative; *Acc*: accuracy; *Se*: sensitivity; *Spe*: specificity; *PPV*: positive predictive value; *NPV*: negative predictive value; sd: standard deviation; *Q*: quantile; LR: logistic regression.

In the results presented above, non-significant differences were verified for *Spe* values between LR1.23 and LR2.17 ($p = 0.89$) and LR1.36 and LR2.35 ($p = 0.96$) models, and *PPV* values between

LR1.23 and LR2.17 ($p = 0.442$) and LR1.36 and LR2.35 ($p = 0.38$) models. A significant difference for $p < 0.05$ was found for *Acc* values between LR1.23 and LR2.17 models ($p = 0.039$), and in all other cases significant differences for $p < 0.01$ were found, with better performance from LR2 model. Considering these results, together with the fact LR2 uses one less explanatory variable, it was therefore considered the most parsimonious alternative, and selected as the best LR model.

4.4 Decision Tree Model

For implementation of the DT model, and as already referred in the methods section, a variation of the traditional DT method has been applied. According to this method, every variable selected at each node has been excluded in the following nodes, so that each variable is used only once, and the DT becomes more understandable from a clinical point of view. In order to select the optimal size DT, the bootstrap resampling method was again applied. To allow subsequent comparisons to the LR model, the same 200 bootstrap samples were used. All data was run through DT with different number of nodes, and OC performance was compared between the different size trees. The DT with least overall error (or maximum accuracy) has been selected. The comparison between the different trees is shown in figure 4.6 and table 4.9.

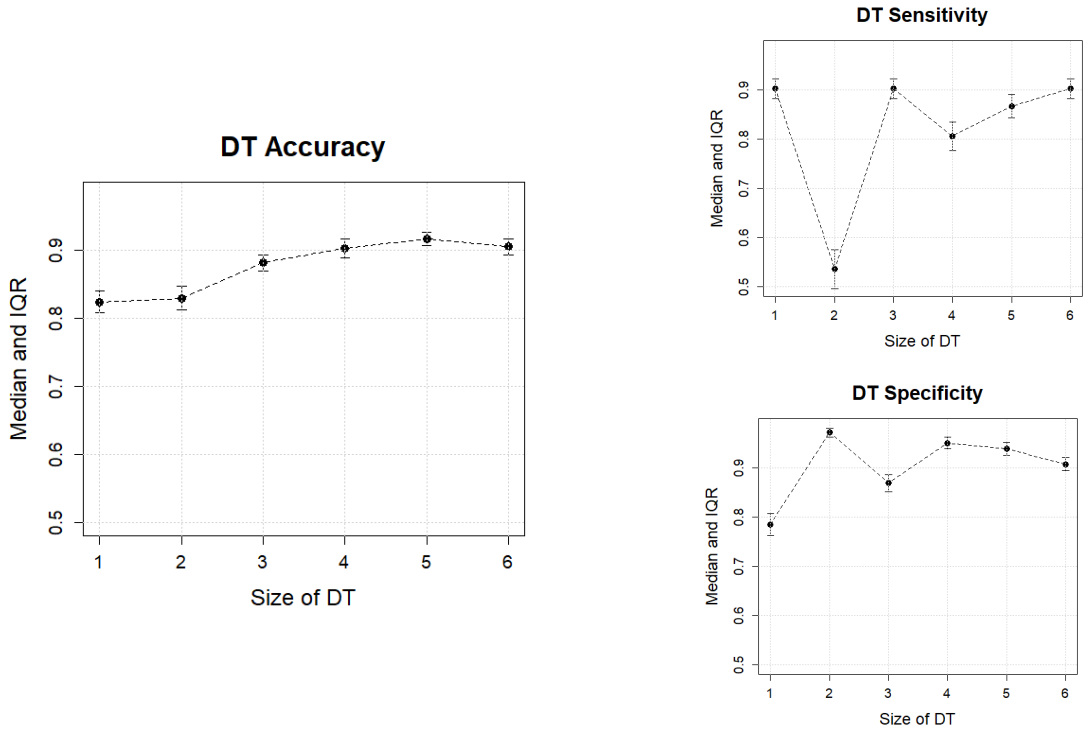


Figure 4.6: Median and interquartile range (IQR) values for DT *Acc*, *Se* and *Spe* with increasing number of variables. The full grown tree comprises the variables LDLc, TG, ApoAI, ApoB, HDLc and TC. DT: decision tree.

Table 4.9: Descriptive statistics for operating characteristics between DT models with increasing number of variables, over 200 bootstrap samples.

Splits	Median	Mean	sd	min	max	$Q_{0.25}$	$Q_{0.75}$
Accuracy							
1	0.82	0.82	0.02	0.75	0.89	0.81	0.84
2	0.83	0.83	0.02	0.77	0.88	0.81	0.85
3	0.88	0.88	0.02	0.82	0.92	0.87	0.89
4	0.90	0.90	0.02	0.85	0.94	0.88	0.91
5	0.92	0.91	0.02	0.86	0.96	0.90	0.92
6	0.90	0.90	0.02	0.85	0.94	0.89	0.92
Sensitivity							
1	0.90	0.90	0.03	0.82	0.98	0.88	0.92
2	0.54	0.54	0.06	0.38	0.69	0.50	0.58
3	0.90	0.90	0.03	0.82	0.98	0.88	0.92
4	0.81	0.80	0.04	0.71	0.94	0.78	0.83
5	0.87	0.87	0.03	0.76	0.95	0.84	0.89
6	0.90	0.90	0.03	0.82	0.98	0.88	0.92
Specificity							
1	0.78	0.78	0.03	0.70	0.87	0.76	0.81
2	0.97	0.97	0.01	0.93	1.00	0.96	0.98
3	0.87	0.87	0.03	0.79	0.93	0.85	0.89
4	0.95	0.95	0.02	0.90	0.99	0.94	0.96
5	0.94	0.94	0.02	0.88	0.98	0.92	0.95
6	0.91	0.90	0.02	0.84	0.95	0.89	0.92

sd: standard deviation; Q: quantile.

As can be seen in the figure presented above, variables are sequentially included until a 6 node tree is grown, consisting of the following parameters: LDLc, TG, ApoAI, ApoB, HDLc and TC. Overall *Acc* increases considerably from the inclusion of the third node, corresponding to ApoAI, and decreases slightly after the inclusion of the sixth node, corresponding to TC, which means the DT is starting to overfit the data at this point. The remaining variable, Lp(a), is never included, since in the construction of the tree, the addition of a seventh node with this variable does not reduce further the entropy of the system. *Se* and *Spe* values were also estimated for each tree. Considering maximum *Acc* value (0.92) is obtained for the DT with five nodes (DT5), this was selected as the best performing tree. A representation of DT5, with respective cutoff values at each node, and showing the tree performance in the original samples, is shown in figure 4.7. A confusion matrix and respective OC, resulting from the application of DT5 criteria to the study sample, is shown in table 4.10

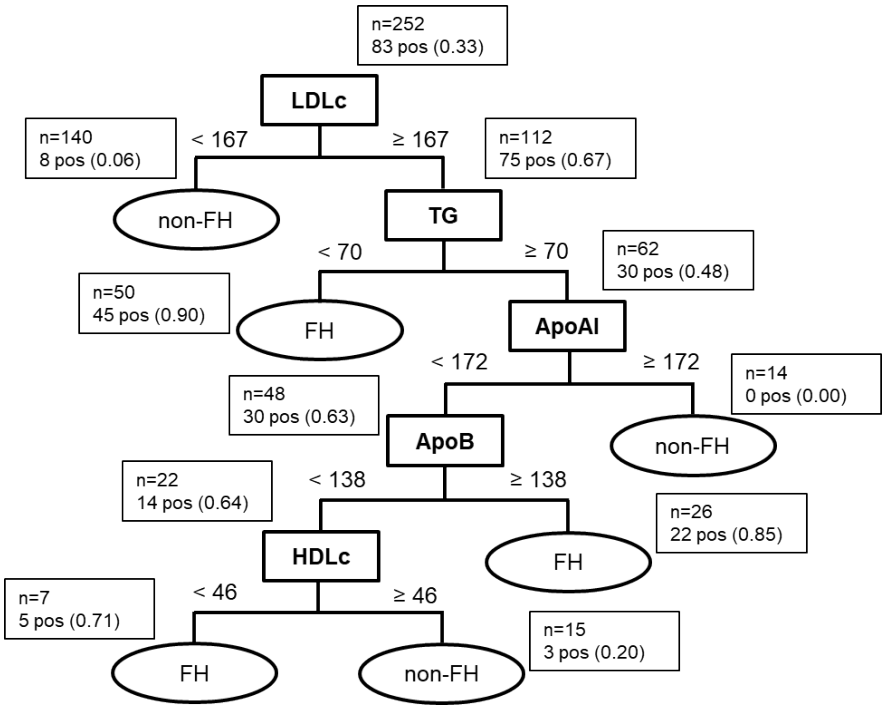


Figure 4.7: Decision tree model with 5 variables. At each node, it is represented the biochemical indicator used to divide the sample, the respective cutoff value, and the way the original sample is divided throughout the tree (n, number and proportion of FH cases). FH: familial hypercholesterolemia; LDLc: low density lipoprotein cholesterol; TG: triglycerides; Apo: apolipoprotein; HDLc: high density lipoprotein cholesterol.

Table 4.10: Confusion matrix and respective operating characteristics for the DT5 model applied to the original sample.

		Molecular test result		
		Positive	Negative	Total
DT method	Positive test	72	11	83
	Negative test	11	158	169
	Total	83	169	252

Op. char.	
Acc	0.91
Se	0.87
Spe	0.93
PPV	0.87
NPV	0.93

DT: decision tree; Acc: accuracy; Se: sensitivity; Spe: specificity; PPV: positive predictive value; NPV: negative predictive value.

4.5 Comparison between Classification Models

A comparative analysis between SB criteria and the best performing LR (LR2) and DT (DT5) models, concerning different operating characteristics, as obtained from the bootstrap samples median and mean results, can be observed in table 4.11 and figure 4.8.

Table 4.11: Descriptive statistics for operating characteristics in SB, LR2 and DT5 models, over 200 bootstrap samples.

	TP	FP	TN	FN	*Acc*	*Se*	*Spe*	*PPV*	*NPV*
SB Criteria									
Median	81.00	65.00	104.00	2.00	0.734	0.976	0.617	0.551	0.982
Mean	80.96	65.30	103.66	2.08	0.733	0.975	0.613	0.554	0.980
sd	7.07	6.68	8.16	1.47	0.027	0.017	0.038	0.038	0.014
min	64	48	86	0	0.663	0.922	0.509	0.466	0.944
max	98	84	123	6	0.798	1.000	0.714	0.653	1.000
$Q_{0.25}$	77.00	60.00	98.00	1.00	0.714	0.964	0.584	0.527	0.970
$Q_{0.75}$	86.00	70.00	109.00	3.00	0.750	0.988	0.640	0.583	0.991
LR2.17									
Median	81.00	29.00	140.00	2.00	0.877	0.978	0.829	0.733	0.986
Mean	81.20	29.04	139.91	1.84	0.877	0.978	0.828	0.737	0.987
sd	7.17	5.15	7.83	1.35	0.021	0.016	0.030	0.042	0.009
min	63	17	119	0	0.829	0.921	0.754	0.627	0.951
max	98	42	159	7	0.929	1.000	0.900	0.839	1.000
$Q_{0.25}$	76.00	25.00	134.00	1.00	0.861	0.967	0.806	0.704	0.980
$Q_{0.75}$	86.00	33.00	145.00	3.00	0.893	0.988	0.850	0.770	0.993
LR2.35									
Median	73.00	13.00	156.00	10.00	0.909	0.882	0.921	0.847	0.940
Mean	73.04	13.49	155.47	10.01	0.907	0.879	0.920	0.844	0.940
sd	6.97	3.63	7.52	2.99	0.019	0.034	0.021	0.039	0.018
min	57	6	136	3	0.849	0.772	0.863	0.735	0.874
max	93	24	176	21	0.956	0.962	0.964	0.935	0.981
$Q_{0.25}$	68.00	11.00	150.00	8.00	0.897	0.855	0.908	0.819	0.927
$Q_{0.75}$	77.00	15.25	160.00	12.00	0.921	0.903	0.934	0.875	0.952
DT5									
Median	72.00	10.50	158.00	11.00	0.917	0.866	0.938	0.874	0.935
Mean	71.85	10.58	158.38	11.20	0.914	0.865	0.937	0.872	0.934
sd	6.70	3.34	7.75	2.95	0.017	0.033	0.020	0.037	0.018
min	56	4	139	3	0.861	0.762	0.877	0.774	0.890
max	89	20	179	19	0.960	0.955	0.978	0.954	0.983
$Q_{0.25}$	67.00	8.00	154.00	9.00	0.905	0.839	0.925	0.852	0.922
$Q_{0.75}$	76.00	13.00	164.00	13.00	0.925	0.886	0.951	0.897	0.946

TP: true positive; FP: false positive; TN: true negative; FN: false negative; *Acc*: accuracy; *Se*: sensitivity; *Spe*: specificity; *PPV*: positive predictive value; *NPV*: negative predictive value; sd: standard deviation; *Q*: quantile; SB: Simon Broome; DT: decision tree; LR: logistic regression.

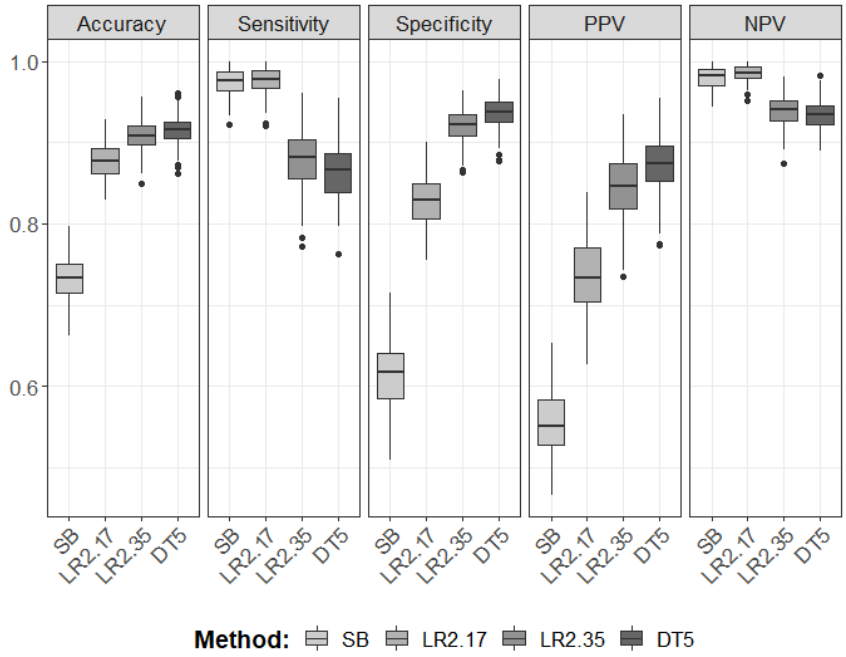


Figure 4.8: Boxplot representation for SB, LR2 and DT5 models among different operating characteristics, in 200 bootstrap samples. SB: Simon Broome; DT: decision tree; LR: logistic regression.

Several differences in OC performance can be seen in the results presented above. Overall Acc is higher in DT5 model, followed by LR2 model with highest cut point ($c = 0.35$). Similar behaviour is found for Spe and PPV. On the other hand, better Se levels are achieved by SB and LR2 model with the lowest cut point ($c = 0.17$), with similar behaviour found for NPV. Using Wilcoxon signed rank test, it is possible to confirm that all differences are significant for $p < 0.01$, except Se values between SB and LR2.17 models ($p = 0.098$), and NPV values between LR2.35 and DT5 models ($p = 0.04$, still significant for $p < 0.05$). Interestingly, while presenting similar NPV and Se levels as SB criteria, LR2.17 model still achieves better Acc, Spe and NPV.

Finally, an attempt was made in order to understand how different methods are classifying individuals in the original sample, by comparison with molecular diagnosis results. A matrix plot was produced for this purpose, where misclassification rate for each method can be visually inspected, as shown in figure 4.9.

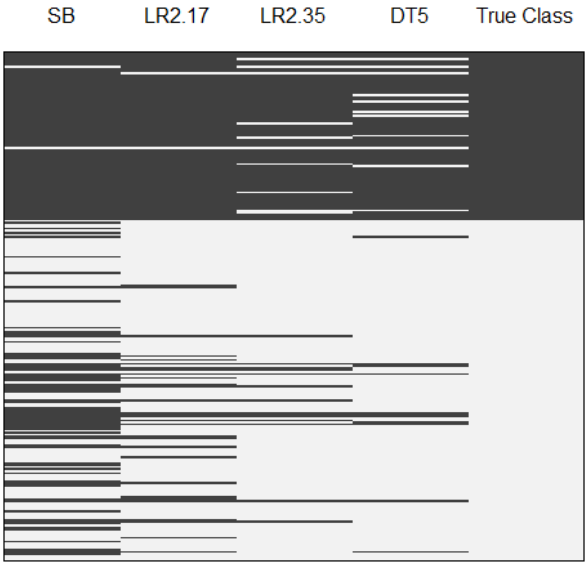


Figure 4.9: Matrix plot representing concordance between SB, LR2 and DT5 models with molecular diagnosis. Positive cases are represented in dark grey and negative cases in light gray. SB: Simon Broome; DT: decision tree; LR: logistic regression.

The observation of this figure seems to confirm that SB criteria present in fact the highest amount of FP cases, as can be seen from the amount of dark grey dashes in subjects that have a true classification of non-FH, as assessed by the molecular study (presented in light grey). In a similar fashion, LR2.35 and DT5 seem to be the methods with lower FP number, but at the same time higher number of FN cases, as can be seen by the amount of light grey dashes in subject that have a true classification of FH (presented in dark grey).

In order to quantify concordance between the several classification methods, two different tables were organized. In table 4.12 its presented the number of subjects, whether FH or non-FH, that are classified in the same way by different combinations of methods. Table 4.13 on the other hand presents the percentage of concordance between the different methods, as well as between these and molecular diagnosis classification.

Table 4.12: Table of concordance between the different classification methods and molecular diagnosis. In each line, it is presented how many, and which methods correctly classify the participant, classified as FH or non-FH by molecular diagnosis.

Nr. correct	Which	FH	non-FH
All		68	99
3	SB + LR2.17 + LR2.35	7	0
	SB + LR2.17 + DT5	1	0
	SB + LR2.35 + DT5	0	3
	LR2.17 + LR2.35 + DT5	0	35
2	SB + DT5	3	1
	LR2.17 + LR2.35	1	3
	LR2.35 + DT5	0	13
1	SB	2	0
	LR2.35	0	1
	DT5	0	7
None		1	7
Total		83	169

* Combinations of methods that have zero counts in FH and non-FH subjects are not represented. FH: familial hypercholesterolemia; SB: Simon Broome; DT: decision tree; LR: logistic regression.

The results from table 4.12 present a very intuitive interpretation of the way different classification methods are in agreement with molecular diagnosis, for the study sample. For example, it can be seen that 68 FH patients are correctly classified by all methods, whether 1 FH patient is incorrectly diagnosed as non-FH by every method. Among non-FH patients, a total of 59 subjects (35+3+13+1+7) are correctly classified by other methods, that are not detected using SB criteria, corroborating the low *Spe* of this method.

Table 4.13: Percentage of concordance between the different classification methods, and true classification as assessed by molecular diagnosis.

	SB	LR2.17	LR2.35	DT5	True Class.
SB		204 (.809)	192 (.762)	188 (.746)	**184 (.730)**
LR2.17			228 (.905)	218 (.865)	**220 (.873)**
LR2.35				230 (.913)	**228 (.905)**
DT5					**230 (.913)**

* The results are presented as: number of concordant subjects (% of concordance). SB: Simon Broome; DT: decision tree; LR: logistic regression.

Finally, overall percentage of concordance was higher between LR2.35 and DT5 classification methods (0.913). These are also the methods that presented higher percentage of concordance with molecular diagnosis (0.905 and 0.913 respectively), in agreement with what was previously observed through bootstrap resampling methods.

Chapter 5

Discussion and Conclusions

This last chapter is dedicated to the discussion of the results and main conclusions of the study. The discussion was divided in different sections, covering several topics: preliminary exploratory analysis of data, individual detailed considerations regarding LR and DT models, and comparative analysis between the different classification models.

5.1 Exploratory data analysis

Before the development of the different classification models, an exploratory analysis of sample data was performed. Through this preliminary analysis, it was possible to identify potentially discrepant values, that may or may not need correction, and to have an idea of the distribution of the several variables between FH and non-FH individuals, as well as the relation between different variables. Important to note, the presented results already refer to the clean dataset, after univariate and bivariate inspection of the different variables was performed, with outlier observations and discordant values verified. Examples of the graphs used for this purpose can be found in Appendix D.

Concerning the sample's characteristics, no significant differences were found for age group ($p = 0.097$) and gender ($p = 0.376$), between FH and non-FH subjects, as assessed by chi-square test. Significant differences for age as a continuous variables were not assessed since this variable has been discretized by rounding to a full year, and the range of ages is quite reduced (2 to 17 years). Other variables, like anthropometric indicators and clinical information have not been included at this point, since this information was not available for all patients, and are to be taken in consideration in a next stage of the project, already underway.

Regarding biochemical indicators, observation of the respective density plots suggests the distribution of these variables differs markedly between FH and non-FH groups, in particular for TC, LDLc, TG and ApoB. Additionally, these do not appear to follow a normal distribution, which has been confirmed by one-sample KS test with Liliefors correction. Significant differences for these values were therefore assessed by the non-parametric two-sample KS test. The KS test was preferred over other non-parametric tests, like the Mann-Whitney-Wilcoxon (MWW) test or the permutation test, because it compares the cumulative distribution function of both groups, and is therefore

sensitive to differences in the shape and median values of both distributions, while the other tests only check for differences between median values. All biochemical variables differed significantly between FH and non-FH patients ($p < 0.01$). As expected, FH patients presented generally higher levels of TC, LDLc, and ApoB, since this disorder predominantly affects LDL metabolic pathway, consequently increasing LDLc and associated components [1]. Non-FH subjects on the other hand presented higher values for TG, HDLc, ApoAI, and Lp(a).

5.2 The LR model

For the development of the LR model, several steps were taken. VIF analysis including all variables revealed the presence of high VIF(> 4), with TC presenting the highest value (VIF=25.30), which suggests elevated colinearity with other variables, and was consequently removed. The remaining variables presented VIF< 4, and were therefore considered as the full model. The fact TC is highly correlated with other variables is understandable, since it represents the sum of the different cholesterol fractions: LDLc, HDLc and remnant cholesterol (VLDL and IDL), indirectly estimated by TG [37]. In particular, TC seems to be highly correlated with LDLc ($r = 0.94$), which also presented very high VIF (VIF=23.04). The fitted model for these variables was obtained by purposeful selection methods [16]. Forward, backward, stepwise forward and stepwise backward methods were applied, and the same model was obtained by all methods, with the same AIC value (176.78). The variables included in the final model were LDLc, TG, ApoAI and Lp(a), ordered by statistical significance. A positive β coefficient for LDLc indicates this variable has a direct relation with the presence of FH, while negative β coefficients for the other variables suggest these have an inverse relation with the presence of FH. Lp(a) was not significant for usual significance levels ($p = 0.059$), but was kept in the model, since use of a conservative p-value until 0.2 is recommended at this stage [14], and most important, the deviance test reveals the model with and without this variable differ significantly ($p = 0.047$). The variables age and gender were included in the full model, although no significant association with the presence of FH has been found, so that they could act as controllers. Since variable selection procedures did not retain these variables, allied with the fact these don't seem to be clinically relevant for FH diagnosis among pediatric patients [4], we decided to exclude them from the final model.

Residual analysis for the LR model revealed the presence of several potentially influential observations, as assessed by different graphical methods: standardized residuals vs linear predictor, outliers identification, leverage, Cook's distance and influential observations plots. Five observations in total were removed, and a new analysis was performed without these cases. As before, TC still presented a high VIF (VIF=18.81), and was excluded from the list of predictor variables in the full model. The final model differed from the previous one, by including only the variables LDLc, TG and ApoAI, all significant for $p < 0.01$. Both models (named LR1 and LR2 respectively), have been therefore considered for subsequent analysis.

For model adjustment, ROC curves were generated for LR1 and LR2, and two different cutoff points were calculated based on distinct methods: Youden index and *minimum* p-value approach. Like referred in the methods section, Youden index maximizes the summation between Se and Spe, corresponding to the point in the ROC curve with highest vertical distance from the $y = x$ diagonal

line. This is therefore the point that maximizes sensitivity, possessing a cut-off value $c = 0.23$ in LR1 and $c = 0.17$ in LR2, c corresponding to the estimated probability of being FH positive. The minimum p-value method on the other hand defines the optimal cutoff point as the one that maximizes standard chi-square statistic. For each cutoff point, a 2x2 contingency table is created, defining $Se(c) = P(\hat{\pi} > c|E = 1)$ and $Spe(c) = P(\hat{\pi} \leq c|E = 0)$. The chosen cut-off point will maximize both of these OC, and will therefore be equivalent to the point with least overall error, or maximum Acc [25]. Following this process, LR1 and LR2 presented cut-off values of $c = 0.36$ and $c = 0.35$, respectively. The OC values, obtained from the respective contingency tables for the selected cutpoints corroborate the premise behind these methods, with higher Se values found with Youden index and higher Acc values with the minimum p-value approach. Because Youden index method privileges Se it is logic that the obtained cutpoints are lower, i.e., a certain patient will be more easily classified as positive than through minimum p-value method. These cutpoints cannot however be compared directly, since they were obtained from different samples.

The same is valid for all the measures referring to the model overall quality: AUC, pseudo-R^2 and GOF measures. Because the variability in LR2 model has been reduced by elimination of the most discrepant observations, seems logic to verify this model presents a better general performance in these indicators. Nevertheless, the quality of adjustment is apparently very good for both models. Concerning GOF measures, HL test rejects the null hypothesis of good model adjustment ($p < 0.05$) for both models, whereas LC test rejects this hypothesis only for LR1. LC test was considered preferable to assess GOF, since HL test requires dividing up the sample in a selected number of groups for application of Pearson chi-squared statistic, with the corresponding test statistic and p-value varying considerably. LC on the other side is based on the weighted sum of smoothed residuals, and does not imply dividing the sample in an arbitrary number of groups [20].

To overcome the limitation of not being able to compare both LR models directly, bootstrap resampling methods were used. 200 bootstrap samples of the complete set of observations ($N = 252$) have been generated, and ran through both LR models, and several central tendency and dispersion measures of the different OC were calculated, and compared through Wilcoxon signed rank test. Results from this analysis have shown that either non-significant differences are observable between both LR models, or that significant differences evidence a better performance for LR2 model. These results, together with the fact LR2 uses one less explanatory variable, led to the decision of keeping LR2 model for further analysis, and exclude LR1.

5.3 The DT model

The other classification method developed in this work, DT model, is used on information theory, and based on entropy reduction measures [27]. As mentioned in the methods section, a modified version of the classical DT was implemented, consisting in the sequential elimination of predictor variables as they are used in each tree node, so that each variable enters the tree only once, hence assuming a structure that typically resembles medical criteria. A major challenge in the implementation of this method was the decision regarding how many nodes should the final tree include for optimal performance, i.e., to provide the maximum amount of information without overfitting the data. In order to select the optimal size DT, the bootstrap resampling method was again uti-

lized, using the same 200 bootstrap samples as the ones used to compare the LR models previously presented, so that subsequent comparisons between models would be possible. Using the original sample, different DT were built, with increasing node number, until all variables entered the model, or until the inclusion of an additional variable did not further reduce the entropy of the system. Following this algorithm, six different DT were obtained, and named according to number of nodes, from DT1 to DT6. Simply put, DT1 consisted of the tree using only LDLc variable, the biochemical indicator that initially provided the highest information gain, with a cutoff point above 167 mg/dL to be classified as FH, and so on, until the full grown tree was obtained, comprising the variables LDLc, TG, ApoAI, ApoB, HDLc and TC, by hierarchical order. As previously mentioned, the remaining variable, Lp(a), was not included, since the addition of a seventh node using this indicator did not reduce the entropy of the system further. The bootstrap samples were then run through the different DT models, and several OC were calculated for comparison purposes: *Acc*, *Se*, and *Spe*. The DT with lowest median misclassification rate, or higher *Acc*, was finally selected from the set of candidate trees. Because overall median *Acc* decreased slightly with the inclusion of the sixth variable, TC, this biochemical indicator was excluded from the DT, and the tree with the remaining 5 variables (DT5), was defined as the final model. This model is represented in figure 4.7 in the results, showing respective cutoff values for each node, and how the original sample is divided throughout the DT. The respective confusion matrix and OC resulting from the application of DT5 criteria to the study sample are also presented.

5.4 Comparison between different classification models

One finding that was found very interesting in the current work is the fact that, either using LR or DT classification methods, the most relevant variables selected to classify the individual as FH or non-FH, are LDLc, TG and ApoAI, by order of importance. While LDLc concentration is directly related to the probability of being FH positive, TG and ApoAI levels are inversely related to the presence of this pathology. These results are also plausible from a biological point of view. First of all, it seems logic to confirm LDLc is the most relevant variable in both models, since this genetic disorder primarily affects LDLc metabolism, causing LDLc levels to increase [1]. TG appear in both models as the second most informative variables, with lower TG levels associated to FH presence. This may be related to the fact that high TG levels are more related to dyslipidemia triggered essentially by environmental factors. Similar results have been previously reported, p.e. in the Welsh population, for which the Dutch Lipid Clinic Network criteria have been modified to take into account that elevated TG levels in a patient with FH phenotype makes it less likely that the patient effectively has FH [38]. Finally, ApoAI arises as the third most relevant variable in both cases, also with lower ApoAI levels related to FH presence. This is expected, since ApoAI is the major constituent apolipoprotein of HDL, which participates in reverse cholesterol transportation. Specifically, it is ApoAI content that determines HDL size and function, including cholesterol removal processes from peripheral cells, interaction with lipids, and responsiveness to specific receptors and proteins [8]. Based on the results of this work, ApoAI content seems to be more determinant than HDL concentration alone to separate FH from non-FH subjects.

The best performing LR model (LR2), using two different cutpoints ($c = 0.17$ and $c = 0.35$)

and the best performing DT model (DT5) were finally compared with each other, as well as with SB criteria, regarding different OC. The same bootstrap samples previously used to validate each of these classification methods were again used for this purpose. *Acc*, *Spe* and PPV median values were higher in DT5 model, followed by LR2 model with $c = 0.35$, with SB criteria performing the worst. This suggests these methods do not only correctly classify patients more often than SB criteria, but also that they possess better ability to exclude negative cases. *Se* and NPV on the other hand were higher in LR2 model with $c = 0.17$ and SB criteria, indicating better ability to retain FH positive cases by these models. However, this seems to be accomplished at the expense of retaining a high number of false positives, which can prove to be costly and inefficient in clinical practice [1]. One of the most relevant results of this comparison is the fact that, while *Se* values between SB and LR2.17 models did not differ significantly ($p = 0.098$), LR2.17 model achieved better performance for all other OC, particularly *Acc*, *Se* and NPV ($p < 0.01$), which to be confirmed would undoubtedly indicate this method as preferable between the two.

Finally, by means of a concordance matrix and tables, it was investigated how different methods are classifying individuals in the original sample, by comparison with molecular diagnosis. The results concerning the original sample are in agreement with the ones obtained through bootstrap resampling analysis. Specifically, LR2.35 and DT5 seem to be the most accurate methods, while LR2.17 and SB criteria seem to present higher *Se*. Also, while LR2.17 and SB criteria have an equivalent performance in detecting FH cases, SB criteria seems to perform worst in ascertaining non-FH cases, as 35 FP cases are retained by this method, that are correctly classified as non-FH by all the other methods, and a total of 59 FP cases are obtained that are correctly classified by at least one of the other methods.

5.5 Conclusions

Several conclusions have been taken from the current work, regarding the performance of different classification methods for FH diagnosis. Higher *Acc*, *Spe* and PPV values were achieved by means of a DT model, or by a LR model using a cutoff point defined by the minimum p-value method. In these cases, overall misclassification rate is lower, as well as false positive retention. Higher *Se* and NPV on the other hand were obtained by means of a LR model using a cutoff point as defined by Youden's index, or using SB criteria, suggesting better ability to retain FH cases by these models. Between these two methods however, the values of the remaining OC differed substantially, with the LR model achieving better performance, which to be validated by additional data would definitely indicate this method as preferable between the two. The poor performance of SB criteria in these OC is due to the use of conservative cutoff values for LDLc and TC, and the high number of false positive cases that are consequently retained by this method can prove to be costly and inefficient in clinical practice.

It seems that, in spite using different approaches, both LR and DT methods are able to divide the sample according to the most relevant biochemical characteristics for FH diagnosis. Specifically, either using LR or DT classification methods, the most relevant variables selected to distinguish FH from non-FH individuals are LDLc, TG and ApoAI, by order of relevance. Compared to other dyslipidemic children, FH individuals seem to possess increased LDLc levels, and relatively lower TG

and ApoAI levels. These findings seem to have biological plausibility, since FH primarily affects LDLc metabolism, resulting in increased LDLc levels, high TG levels may be more related to dyslipidemia triggered essentially by environmental factors, and high ApoAI levels are indicative of an efficient reverse cholesterol transportation mechanism, which may also be diminished in FH patients.

Between each other, LR and DT models possess distinct advantages and disadvantages. In LR models, cutoff values can be adjusted according to different methods, to better suit the purpose of the decider. Different cutoff values can be taken from the same LR model and compared, or used together for classification purposes. In the current study, Youden index and minimum p-value methods were used to define cutoff values. While Youden index produces a relatively lower cutoff value, and maximizes Se values, the minimum p-value approach produces a higher cutoff value, and maximizes Acc levels. Other statistical procedures inherent to LR models, like residuals analysis, VIF and model performance measures, such as AUC, pseudo-R^2 and GOF measures contribute to make this method more robust. In the current study p.e., potentially influential observations were signalled, and the model built without these observations was finally validated as the most efficient, through bootstrap methods. Also, through VIF analysis, TC was detected to be highly correlated with other biochemical variables (which is logic considering it represents the sum of the different cholesterol fractions), and removed from the pool of candidate predictor variables.

The DT model on the other hand, has the advantage of providing a very visual classification tool, with specific cutoff values for each biochemical variable. The DT model in this study was further modified to sequentially eliminate predictor variables as they are used, so that each variable enters the tree only once, hence assuming a structure that typically resembles medical criteria, and can therefore be easily used in clinical practice. Some of the features presented in one of the models can also be used to improve the predictive ability of the other model. For example, residuals analysis and VIF can be used as a preliminary step before the development of a DT. The DT model can also be adapted to divide the sample according to a predetermined Se or Acc level, and distinct DT can potentially be built from the same sample according to different criteria, mimicking LR models with different cutpoints. Future work is being prepared in this sense.

References

[1] O. Najam and K. K. Ray, "Familial hypercholesterolemia: a review of the natural history, diagnosis, and management," *Cardiology and therapy*, vol. 4, no. 1, pp. 25–38, 2015.

[2] M. Sharifi, M. Futema, D. Nair, and S. E. Humphries, "Genetic architecture of familial hypercholesterolaemia," *Current cardiology reports*, vol. 19, no. 5, p. 44, 2017.

[3] V. E. Bouhairie and A. C. Goldberg, "Familial hypercholesterolemia," *Cardiology clinics*, vol. 33, no. 2, pp. 169–179, 2015.

[4] A. C. Martin, S. S. Gidding, A. Wiegman, and G. F. Watts, "Knowns and unknowns in the care of pediatric familial hypercholesterolemia," *Journal of lipid research*, vol. 58, no. 9, pp. 1765–1776, 2017.

[5] M. A. Austin, C. M. Hutter, R. L. Zimmern, and S. E. Humphries, "Genetic causes of monogenic heterozygous familial hypercholesterolemia: a huge prevalence review," *American journal of epidemiology*, vol. 160, no. 5, pp. 407–420, 2004.

[6] R. of fatal coronary heart disease in familial hypercholesterolaemia, "Scientific steering committee on behalf of the simon broome register group," *BMJ*, vol. 303, no. 6807, pp. 893–6, 1991.

[7] B. G. Nordestgaard, M. J. Chapman, S. E. Humphries, H. N. Ginsberg, L. Masana, O. S. Descamps, O. Wiklund, R. A. Hegele, F. J. Raal, J. C. Defesche, *et al.*, "Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the european atherosclerosis society," *European heart journal*, vol. 34, no. 45, pp. 3478–3490, 2013.

[8] K. R. Feingold and C. Grunfeld, "Introduction to lipids and lipoproteins," in *Endotext [Internet]*, MDText. com, Inc., 2018.

[9] L. R. Engelking, *Textbook of veterinary physiological chemistry.* Academic Press, 2014.

[10] M. M. Hussain, "Intestinal lipid absorption and lipoprotein formation," *Current opinion in lipidology*, vol. 25, no. 3, p. 200, 2014.

[11] C. Wiener, A. S. Fauci, E. Braunwald, D. L. Kasper, S. L. Hauser, D. L. Longo, J. L. Jameson, and J. Loscalzo, *Harrison's principles of internal medicine, self-assessment and board review.* McGraw Hill Professional, 2008.

[12] M. A. Iacocca, J. R. Chora, A. Carrié, T. Freiberger, S. E. Leigh, J. C. Defesche, C. L. Kurtz, M. T. DiStefano, R. D. Santos, S. E. Humphries, *et al.*, "Clinvar database of global familial hypercholesterolemia-associated dna variants," *Human mutation*, vol. 39, no. 11, pp. 1631–1640, 2018.

[13] A. Medeiros, A. Alves, V. Francisco, M. Bourbon, *et al.*, "Update of the portuguese familial hypercholesterolaemia study," *Atherosclerosis*, vol. 212, no. 2, pp. 553–558, 2010.

[14] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.

[15] A. J. Dobson and A. Barnett, *An introduction to generalized linear models*. Chapman and Hall/CRC, 2008.

[16] Z. Bursac, C. H. Gauss, D. K. Williams, and D. W. Hosmer, "Purposeful selection of variables in logistic regression," *Source code for biology and medicine*, vol. 3, no. 1, p. 17, 2008.

[17] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: Logistic regression," *Perspectives in clinical research*, vol. 8, no. 3, p. 148, 2017.

[18] S. K. Sarkar, H. Midi, and S. Rana, "Detection of outliers and influential observations in binary logistic regression: An empirical study," *Journal of Applied Sciences*, vol. 11, no. 1, pp. 26–35, 2011.

[19] P. D. Allison, "Measures of fit for logistic regression," in *SAS Global Forum, Washington, DC*, 2014.

[20] D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow, "A comparison of goodness-of-fit tests for the logistic regression model," *Statistics in medicine*, vol. 16, no. 9, pp. 965–980, 1997.

[21] C. Silvia-Fortes, "Testes de diagnósstico e curvas roc," in *Bioestatística e Qualidade na Saúde* (G. Cunha, M. Eiras, and N. Teixeira, eds.), 2011.

[22] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[23] W. Zhu, N. Zeng, N. Wang, *et al.*, "Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations," *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, vol. 19, p. 67, 2010.

[24] F. Habibzadeh, P. Habibzadeh, and M. Yadollahie, "On determining the most appropriate test cut-off value: the case of tests with continuous results," *Biochemia medica: Biochemia medica*, vol. 26, no. 3, pp. 297–307, 2016.

[25] I. Unal, "Defining an optimal cut-point value in roc analysis: an alternative approach," *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[26] M. Greiner, D. Pfeiffer, and R. Smith, "Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests," *Preventive veterinary medicine*, vol. 45, no. 1-2, pp. 23–41, 2000.

[27] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[28] M. J. Berry and G. Linoff, *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.

[29] M. Antunes, *Prospecção de dados: princípios e métodos*. Centro de Estatística e Aplicações da U.L., 2015.

[30] T. Hesterberg, D. S. Moore, S. Monaghan, A. Clipson, and R. Epstein, "Bootstrap methods and permutation tests," *Introduction to the Practice of Statistics*, vol. 5, pp. 1–70, 2005.

[31] K. Singh and M. Xie, "Bootstrap: a statistical method," *Unpublished manuscript, Rutgers University, USA. Retrieved from http://www. stat. rutgers. edu/home/mxie/RCPapers/bootstrap. pdf*, 2008.

[32] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[33] C. J. Jolliffe, I. Janssen, *et al.*, "Distribution of lipoproteins by age and gender in adolescents," *Circulation*, vol. 114, no. 10, pp. 1056–1062, 2006.

[34] A. Benito-Vicente, A. C. Alves, A. Etxebarria, A. M. Medeiros, C. Martin, and M. Bourbon, "The importance of an integrated analysis of clinical, molecular, and functional data for the genetic diagnosis of familial hypercholesterolemia," *Genetics in Medicine*, vol. 17, no. 12, p. 980, 2015.

[35] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2018.

[36] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[37] B. G. Nordestgaard, A. Langsted, S. Mora, G. Kolovou, H. Baum, E. Bruckert, G. F. Watts, G. Sypniewska, O. Wiklund, J. Borén, *et al.*, "Fasting is not routinely required for determination of a lipid profile: clinical and laboratory implications including flagging at desirable concentration cut-points—a joint consensus statement from the european atherosclerosis society and european federation of clinical chemistry and laboratory medicine," *European heart journal*, vol. 37, no. 25, pp. 1944–1958, 2016.

[38] K. Haralambos, S. Whatley, R. Edwards, R. Gingell, D. Townsend, P. Ashfield-Watt, P. Lansberg, D. Datta, and I. McDowell, "Clinical experience of scoring criteria for familial hypercholesterolaemia (fh) genetic testing in wales," *Atherosclerosis*, vol. 240, no. 1, pp. 190–196, 2015.

Appendices

Appendix A: Informed consent for participation in the Portuguese FH Study

DECLARAÇÃO DE CONSENTIMENTO

Considerando a "Declaração de Helsínquia" da Associação Médica Mundial
(Helsínquia 1964; Tóquio 1975; Veneza 1983; Hong Kong 1989; Somerset West 1996 e Edimburgo 2000)

Estudo Português de Hipercolesterolemia Familiar

ENQUADRAMENTO DO ESTUDO

Para que se possa compreender e atuar a nível da prevenção das doenças cardiovasculares é necessário a realização de estudos de investigação nesta área. Alguns dos fatores de risco mais importantes para doenças cardiovasculares são a hipertensão arterial, a diabetes, o tabagismo e a dislipidemia. Algumas dislipidemias têm origem genética, nomeadamente a Hipercolesterolemia Familiar (FH). A identificação precoce da causa genética da hipercolesterolemia permite obter o diagnóstico correto da dislipidemia fundamentando a introdução atempada de medidas terapêuticas e aconselhamento de estilos de vida de modo a melhorar o prognóstico do doente. Doentes com Hipercolesterolemia Familiar corretamente identificados e tratados podem ver a sua esperança de vida aumentada em 20/30 anos.

O QUE É O ESTUDO PORTUGUÊS DE HIPERCOLESTEROLEMIA FAMILIAR?

O Estudo Português de Hipercolesterolemia Familiar é um estudo de investigação, cujo objetivo principal é determinar a causa genética da hipercolesterolemia em indivíduos com o diagnóstico clínico de Hipercolesterolemia Familiar, assim como aumentar o conhecimento sobre os mecanismos do desenvolvimento da doença cardiovascular nas pessoas afetadas. Este estudo está autorizado pela Comissão de Ética do Instituto Nacional de Saúde Doutor Ricardo Jorge e pela Comissão Nacional de Proteção de Dados.

O QUE TEREI DE FAZER PARA PARTICIPAR E QUE ANÁLISES IRÃO SER REALIZADAS?

No âmbito deste estudo será realizada uma colheita de sangue, cerca de 15-20 ml no total, para a determinação de parâmetros laboratoriais de relevância e extração de DNA/RNA para estudos moleculares. Será ainda preenchido um questionário onde serão pedidas informações sobre alguns dos seus dados pessoais e elementos da sua história clínica importantes para este estudo.
As análises efetuadas serão gratuitas, porém, não haverá qualquer tipo de compensação financeira pela sua participação ou deslocação.
Estes estudos são confidenciais e o anonimato dos participantes está assegurado de acordo com as normas éticas dos estudos genéticos: os dados referentes à sua identidade não constarão das bases de dados do estudo uma vez que esta informação estará codificada. Só o investigador responsável e o seu assistente é que conhecem a chave de descodificação do participante. As amostras e informações recolhidas serão utilizadas apenas para o presente trabalho de investigação cujo objetivo está descrito acima; as amostras serão guardadas pelo tempo que este estudo estiver em curso, seja para encontrar a causa da dislipidemia, ou para estudar a relação bioquímica/genética, de forma a aumentar o conhecimento dos mecanismos da doença.

IREI TER ACESSO AOS RESULTADOS DAS ANÁLISES PARA AS QUAIS TIREI SANGUE?

Como participante no estudo irá ter acesso aos seus dados clínicos e moleculares, podendo exigir ser retirado deste estudo se assim o desejar. Todos os resultados obtidos serão enviados para o seu médico assistente. Os resultados das análises bioquímicas serão enviados no prazo de 2 semanas e os resultados moleculares serão enviados no prazo máximo de 12 meses.

IMPORTÂNCIA DOS REGISTOS NACIONAIS E INTERNACIONAIS

Os registos nacionais e internacionais de doenças específicas são importantes para a observação da realidade nacional, europeia ou internacional, permitindo o aumento do conhecimento específico de determinada doença, neste caso da Hipercolesterolemia Familiar, em termos da sua prevalência, distribuição e controlo. Este conhecimento irá permitir desenhar estratégias nacionais e/ou internacionais para melhorar a qualidade e esperança de vida destes doentes nomeadamente através da publicação de orientações sobre a identificação e tratamento precoce destes doentes de forma a reduzir o seu elevado risco cardiovascular.
Os dados pessoais como nome, morada e outros contactos, não serão inseridos nestes registos, sendo por esta razão chamados de registos anonimizados.

Instituto Nacional de Saúde Doutor Ricardo Jorge

DECLARAÇÃO DE CONSENTIMENTO

Considerando a "Declaração de Helsínquia" da Associação Médica Mundial
(Helsínquia 1964; Tóquio 1975; Veneza 1983; Hong Kong 1989; Somerset West 1996 e Edimburgo 2000)

Estudo Português de Hipercolesterolemia Familiar

Eu, abaixo-assinado, (nome completo do participante do estudo)
_____,

compreendi a explicação que me foi fornecida, por escrito e verbalmente, acerca da investigação que se tenciona realizar, bem como do estudo em que irei participar. Foi-me dada oportunidade de fazer as perguntas que julguei necessárias, e para todas obtive resposta satisfatória. Tomei conhecimento de que, de acordo com as recomendações da Declaração de Helsínquia, a informação ou explicação que me foi prestada versou os objetivos, os métodos, os benefícios previstos, os riscos potenciais, o eventual desconforto e quais os resultados que me serão comunicados e de que modo. Além disso, foi-me afirmado que tenho o direito de recusar a todo o tempo a minha participação no estudo, sem que isso possa ter como efeito qualquer prejuízo na assistência que me é prestada. Foi-me dado todo o tempo de que necessitei para refletir sobre esta proposta de participação. Nestas circunstâncias:

	Sim	Não
Declaro que fui informado dos objetivos deste estudo e aceito participar nele.		
Autorizo o **tratamento anonimizado e automatizado** dos meus dados pessoais		
Autorizo a **inserção** dos meus dados **anonimizados** num **Registo Nacional de Hipercolesterolemia Familiar**		
Autorizo a **inserção** dos meus dados **anonimizados** num **Registo Internacional de Hipercolesterolemia Familiar**		
Autorizo a **publicação** dos **resultados** obtidos **anonimizados** em artigos em revistas nacionais e internacionais		
Caso não seja possível **contactar diretamente** com o meu médico, autorizo o contacto pela equipa de investigação para o seguimento deste estudo		
Compreendo que este é um **estudo de investigação** cujo tempo de resposta dependerá das condições de financiamento atuais, mas o meu médico assistente receberá um resultado no prazo máximo de 12 meses		

Localidade: _____ Data:_____/_____/_____

Nome do Participante: _____

Assinatura: _____

(Em caso do caso-índex ser menor de 18 anos, os pais devem assinar para dar o seu consentimento)

Nome do Médico Assistente ou do Investigador: _____

Assinatura: _____

Investigadora Responsável do Estudo Português de Hipercolesterolemia Familiar
Doutora Mafalda Bourbon. Tel: 217 508 130 / 217 508 126; Email: mafalda.bourbon@insa.min-saude.pt

NOTA: DEVE SER TIRADA UMA COPIA DA DECLARAÇÃO DE CONSENTIMENTO E ENTREGAR AO PARTICIPANTE OU IMPRIMIR E ASSINAR EM DUPLICADO

Appendix B: Patient data form concerning clinical and biochemical variables

Estudo Português de Hipercolesterolemia Familiar

Adaptado do "Simon Broome Heart Research Trust"
Confidencial

Número do processo: _____
Número da família: _____
(A preencher pelo INSA)

Identificação do caso-índex

Nome completo: _____

Morada: _____

_____

Telefone: _____ Email: _____

Data de Nasc.: _____

Estado Civil: _____

Natural de: _____ Naturalidade dos pais: Pai _____ / Mãe _____

Sexo: Masculino ☐ Feminino ☐

Origem étnica (assinalar a opção adequada):

☐ Caucasiano ☐ Asiático ☐ Africano ☐ Outro

Condições de colheita e envio das amostras

Devem ser feitas as seguintes colheitas em jejum:

CASO-INDEX:

Adultos:
7.5 mL de sangue em tubo de soro com gel separador - após centrifugação (3500rpm/10minutos)
~11 mL de sangue total (4x tubos de EDTA 2.7mL)
Crianças:
5 mL de sangue em tubo de soro com gel separador - após centrifugação (3500rpm/10minutos)
~11 mL de sangue total (4x tubos de EDTA 2.7mL)

FAMILIARES (adultos e crianças):
5 mL de sangue em tubo de soro com gel separador - após centrifugação (3500rpm/10minutos)
~8 mL de sangue total (3x tubos de EDTA 2.7mL)

Estas amostras devem ser enviadas em **correio azul**, em envelope almofadado ou numa caixa, bem envolvidas em algodão, acompanhadas por este formulário.
O tempo máximo entre a colheita das amostras e a sua análise não deverá ultrapassar os **dois dias**, o que implica que a amostra seja enviada logo após a colheita. Esta condição é de extrema importância para este estudo. Para qualquer informação complementar contactar: Doutora Mafalda Bourbon, Tel: 217 508 130 / 217 508 126.

O formulário clínico deve ser sempre acompanhado pela Declaração de Consentimento

REPÚBLICA PORTUGUESA
SAÚDE

SNS SERVIÇO NACIONAL DE SAÚDE

Instituto Nacional de Saúde Doutor Ricardo Jorge

Estudo Português de Hipercolesterolemia Familiar

Inquérito adaptado do "Simon Broome Heart Research Trust"
Confidencial

Número do processo:	
Número da família	
(A preencher pelo INSA)	

Razão primária de inclusão no estudo:
(assinalar a opção adequada)

- ☐ Parente afectado
- ☐ Parente com doença coronária crónica
- ☐ Caso-index tem doença coronária crónica
- ☐ Caso-index sofre de outras doenças vasculares
- ☐ Sinais físicos
- ☐ Screening
- ☐ Outros

Critérios para admissão no estudo (segundo os critérios abaixo mencionados):

- ☐ Hipercolesterolemia familiar confirmada
- ☐ Hipercolesterolemia familiar possível

Critérios para diagnóstico:

Hipercolesterolemia familiar confirmada é definida como:

Crianças menores de 16 anos: Colesterol total acima de 260 mg/dL (6,7 mmol/L) ou LDL colesterol acima de 155 mg/dL (4.0 mmol/L)
Adultos: Colesterol total acima de 290 mg/dL (7,5 mmol/L) ou LDL colesterol acima de 190 mg/dL (4.9 mmol/L).

e

(a) Xantomas nos tendões no caso-index ou parente (pais, filhos avós, irmãos, tios)

ou

(b) Evidência genética de mutação nos genes *LDLR*, *APOB* ou *PCSK9*.

Hipercolesterolemia familiar possível é definida como:

(a)

e

(b) História familiar de enfarte do miocárdio antes dos 50 anos em avós e tios ou antes dos 60 anos nos pais, irmãos e filhos

ou

(c) História familiar de nível elevado de colesterol nos pais, irmãos ou filhos; ou colesterol total acima de 290 mg/dL (7,5 mmol/L) nos avós e/ou tios.

Identificação do Médico Assistente:

Nome: _____

Telefone: _____ Email: _____

Hospital: _____

Serviço: _____

Morada: _____

_____

_____

História médica do caso-index

Valores antes do tratamento

Colesterol mg/dl	LDL. mg/dl	HDL mg/dl	TG mg/dl	ApoB mg/dl	ApoAI mg/dl	Lp(a) mg/dl

Xantomas nos tendões:
(assinale se o caso index alguma vez apresentou xantomas)

	Presente	Ausente		Presente	Ausente
Dorso das mãos	☐	☐	Pretibiais	☐	☐
Cotovelos	☐	☐	Dorso dos pés	☐	☐
			Tendão de Achilles	☐	☐

Olhos

	Presente	Ausente
Lipaemia retinalis	☐	☐
Arco corneano	☐	☐
Xantelasma	☐	☐

Assinalar se o caso-index tem ou teve algumas das seguintes situações:

	Confirmado	Possível	
Angina	☐	☐	Idade de início (anos): _____
Enfarte do miocárdio	☐	☐	Idade do 1º enfarte: _____
CABG	☐	☐	Se sim, quando? _____
Angioplastia	☐	☐	Se sim, quando? _____
Outras doenças vasculares	☐	☐	Se sim, quando? _____
Hipertensão	☐	☐	
A.V.C.	☐	☐	Idade do 1ºA.V.C. _____
A.I.T.	☐	☐	
Claudicação	☐	☐	
Pancreatite	☐	☐	
Doença da tiroide	☐	☐	
Doença renal	☐	☐	
Diabetes	☐	☐	Idade de início (anos): _____
Tratamento da Diabetes:	Insulina	☐	
	Oral	☐	
	Insulina e oral	☐	
	Dieta	☐	

Informação médica da consulta mais recente

Data: _____ Terapêutica: Sim ☐ Não ☐

Colesterol (mg/dl)	LDL (mg/dl)	HDL (mg/dl)	TG (mg/dl)	ApoB (mg/dl)	ApoAI (mg/dl)	Lp(a) (mg/dl)

Altura: _____ (m)
Peso: _____ (kg) IMC: _____ (kg/m^2)

Pressão arterial: Sístole _____ Diástole _____

Álcool: Número de unidades por semana: _____
(1 unidade = 1 cerveja ou um copo de vinho)

Fumador? Sim ☐ Quantos cigarros por dia? _____
Não ☐ Se é um ex-fumador há quantos anos deixou de fumar?_____

Faz exercício? Sim ☐ Que tipo de exercício? _____ Quantas vezes por semana? _____
Não ☐

História familiar do caso-índex (preenchimento obrigatório)

Parentesco	Colesterol elevado	Presença de xantomas	Triglicerideos elevado	Idade do 1º enfarte do miocardio	Vivo (V) Morto (M) Desc.(?)	Idade presente ou na morte	Causa de morte
Pai Nome:	mg/dl		mg/dl				
Mãe Nome:	mg/dl		mg/dl				
Irmão M/F Nome:	mg/dl		mg/dl				
Irmão M/F Nome:	mg/dl		mg/dl				
Irmão M/F Nome:	mg/dl		mg/dl				
Filho M/F Nome:	mg/dl		mg/dl				
Filho M/F Nome:	mg/dl		mg/dl				
Filho M/F Nome:	mg/dl		mg/dl				
Cônjuge Nome:	mg/dl		mg/dl				

Preencher com SIM e NÃO quando os valores não são conhecidos

NOTA: Preencher também o Anexo A caso envie amostras de familiares

Árvore genealógica

Tratamento:

Dieta	☐	Estanois e Fitoesterois	☐
LDL Aferese	☐	Medicamentos	☐

Terapia actual para baixar os lípidos

Estatinas	☐ Qual?_____ Dosagem _____ Data do início do tratamento _____
Resinas	☐ Qual?_____ Dosagem _____ Data do início do tratamento _____
Fibratos	☐ Qual?_____ Dosagem _____ Data do início do tratamento _____
Ácido nicotínico ou derivado	☐ Qual?_____ Dosagem _____ Data do início do tratamento _____
Inibidor da absorção intestinal do colesterol	☐ Qual?_____ Dosagem _____ Data do início do tratamento _____
Outro	☐ Qual?_____ Dosagem _____ Data do início do tratamento _____

Outras observações:

Appendix C: R script used to implement the modified DT model

```
1
2  ##########################################
3  ## IMPLEMENTATION OF SEQUENTIAL DT MODEL:##
4  ##########################################
5
6
7  # I. Load the dataset (data cleaning and exploratory analysis already performed):
8
9  completa <- read.table("completa.csv", header = TRUE, sep = ";", dec = ",", fill = TRUE,
       stringsAsFactors = FALSE)
10 summary (completa)
11
12
13 # II. ENTROPY ANALYSIS:
14
15 # 1. Function to calculate the entropy of the system:
16
17 system_entropy <- function (x){
18   p_positives <- nrow(subset(x, Final_Classification=="Positive_htz"))/NROW(x$Final_
       Classification)
19   p_negatives <- nrow(subset(x, Final_Classification=="Negative"))/NROW(x$Final_
       Classification)
20   S_Entrophy <- (- p_positives * (log2(p_positives))) + (- p_negatives * (log2(p_negatives))
       )
21 }
22 #test <- system_entropy(completa)
23
24
25 # 2. Function to select each possible candidate cutoff value, for a certain numerical
       variable:
26
27 # For each variable:
28 # step 1: For a determined variable, sort values in ascending order; do not consider NA
       values;
29 # step 2: Eliminate consecutive values that have the same classification;
30 # step 3: Consider only unique values;
31 # step 4: Consider as possible cutoff points the mean of consecutive values that correspond
       to a different outcome;
32
33 pontos_de_corte <- function (x){
34   values <- sort(x)
35   values <- subset(values, (!is.na(values)))
36   values <- unique(values)
37   n <- length (values)
38   pontos_corte <- c()
39   for (i in 1:(n-1)){
40     pontos_corte[i] <- (values[i]+values[i+1])/2
41   }
42   return (pontos_corte)
43 }
44
45
46 # 3. Function to calculate de information gain of each of the candidate cutoff values,
47
48 # The function receives two arguments, corresponding to the vector of the numerical variable
49 # and respective dataframe that keeps all variables;
```

```r
# step 1: Calculates the proportion of FH+ and FH- individuals above and below each possible cutoff value;
# step 2: Calcultates the entropy and information gain for each of the candidate cutpoints;

Ent_fun <- function(x,z) {
  p_pos_acima_pc <- c()
  p_neg_acima_pc <- c()
  p_pos_abaixo_pc <- c()
  p_neg_abaixo_pc <- c()
  p_acima_pc <- c()
  p_abaixo_pc <- c()
  Ent_sim_pc <- c()
  Ent_nao_pc <- c()
  Ganho_pc <- c()
  y <- pontos_de_corte(x)
  for (i in seq_along(y)){
    p_pos_acima_pc[i] <- nrow(subset(z, Final_Classification=="Positive_htz"
                                     & x > y[i]))/nrow(subset(z, x > y[i]))
    p_neg_acima_pc[i] <- nrow(subset(z, Final_Classification=="Negative"
                                     & x > y[i]))/nrow(subset(z, x > y[i]))
    p_pos_abaixo_pc[i] <- nrow(subset(z, Final_Classification=="Positive_htz"
                                      & x < y[i]))/nrow(subset(z, x < y[i]))
    p_neg_abaixo_pc[i] <- nrow(subset(z, Final_Classification=="Negative"
                                      & x < y[i]))/nrow(subset(z, x < y[i]))
    p_acima_pc[i] <- nrow(subset(z, x > y[i]))/NROW(z)
    p_abaixo_pc[i] <- nrow(subset(z, x < y[i]))/NROW(z)
    Ent_sim_pc[i] <- (-p_pos_acima_pc[i]*log2(p_pos_acima_pc[i])) + (-p_neg_acima_pc[i]*log2(p_neg_acima_pc[i]))
    Ent_nao_pc[i] <- (-p_pos_abaixo_pc[i]*log2(p_pos_abaixo_pc[i])) + (-p_neg_abaixo_pc[i]*log2(p_neg_abaixo_pc[i]))
    Ganho_pc[i] <- system_entropy(z) - (Ent_sim_pc[i]*p_acima_pc[i] + Ent_nao_pc[i]*p_abaixo_pc[i])
  }
  return(Ganho_pc)
}


# 4. Loop to calculate as optimal cutoff value the one with higher information gain, for each numerical variable,
# and select the variable with higher information gain:

# List of initial candidate variables: TC, LDL, HDL, TG, ApoAI, ApoB, Lpa;

lista.var <- c("Lipids2_TC", "Lipids2_LDLc", "Lipids2_HDLc", "Lipids2_TG",
               "Lipids2_ApoAI", "Lipids2_ApoB", "Lipids2_Lpa")
variaveis <- completa[lista.var]
Ganho_pc <- list()
for (k in 1:length(variaveis)){
    Ganho_pc[[k]] <- Ent_fun(variaveis[[k]], completa)
}
names(Ganho_pc) <- c("Lipids2_TC", "Lipids2_LDLc", "Lipids2_HDLc", "Lipids2_TG")
posicao_PC_opt <- lapply(Ganho_pc, function(x) which.max(x)); posicao_PC_opt
Ganho_pc <- lapply(Ganho_pc, function(x) (x[!is.na(x)])); Ganho_pc
Ganho_max <- lapply(Ganho_pc, function(x) max(x)); Ganho_max
# corte_opt <- lapply(variaveis, function(x) pontos_de_corte(x)); corte_opt
# corte_opt2 <- corte_opt[posicao_PC_opt]
pontos_de_corte(completa$Lipids2_LDLc)[71]
# proporcao de positivos e negativos:
```

```
103 p_positives <- nrow(subset(completa, Final_Classification=="Positive_htz"))/NROW(completa$
        Final_Classification); p_positives
104 p_negatives <- nrow(subset(completa, Final_Classification=="Negative"))/NROW(completa$Final_
        Classification); p_negatives
105
106
107 # First node division: LDL above 167 mg/ dL
108 # Divide the sample according to this variable and cutoff value:
109
110 grupo_1 <- subset(completa, Lipids2_LDLc > 167); nrow(grupo_1)
111 grupo_2 <- subset(completa, Lipids2_LDLc < 167); nrow(grupo_2)
112
113
114 # 2nd level: Exclude this variable and repeat the procedure for the two groups, excluding
        the variable LDL;
115 # Select the variable and cutoff value with highest value of the two groups;
116 # Divide the sample accordingly, and repeat the procedure for the resulting groups,
        excluding another variable;
117 # Repeat the process until all variables have been used, or inclusion of another variable
        does not
118 # reduce the entropy of the system.
```

DT_script_JA

Appendix D: Examples of plots for exploratory data analysis

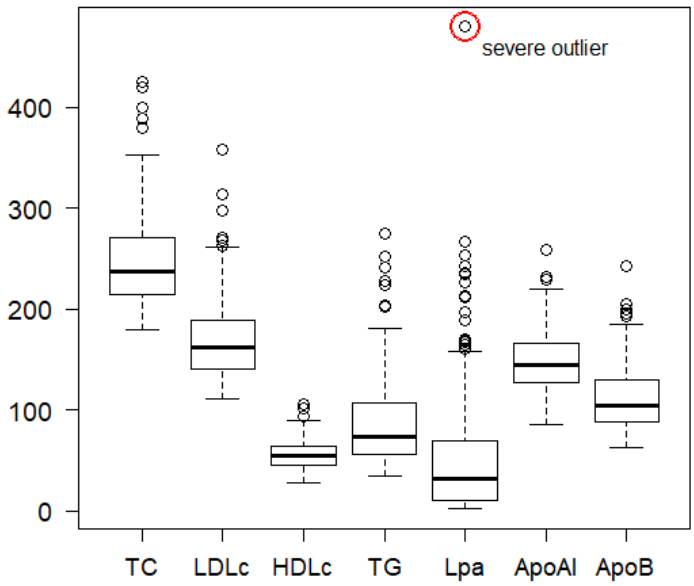


Figure 1: Boxplots of different biochemical variables.

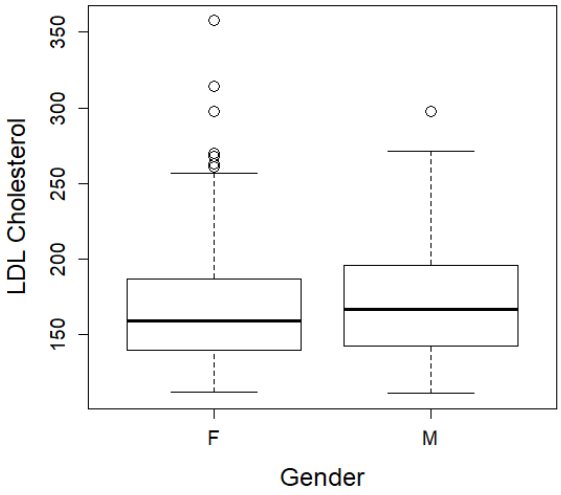


Figure 2: Boxplot of LDL cholesterol values according to gender.

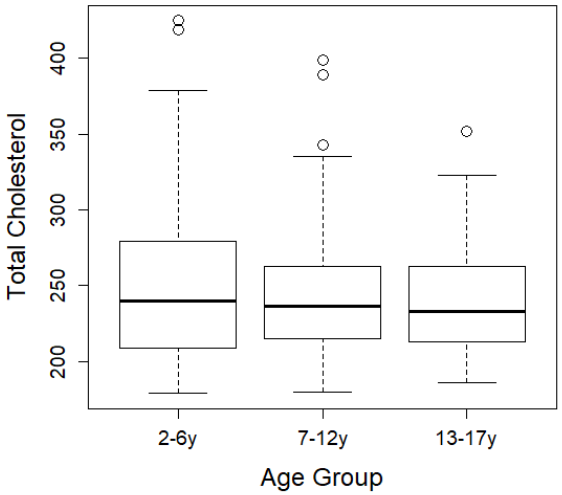


Figure 3: Boxplot of total cholesterol values according to age group.

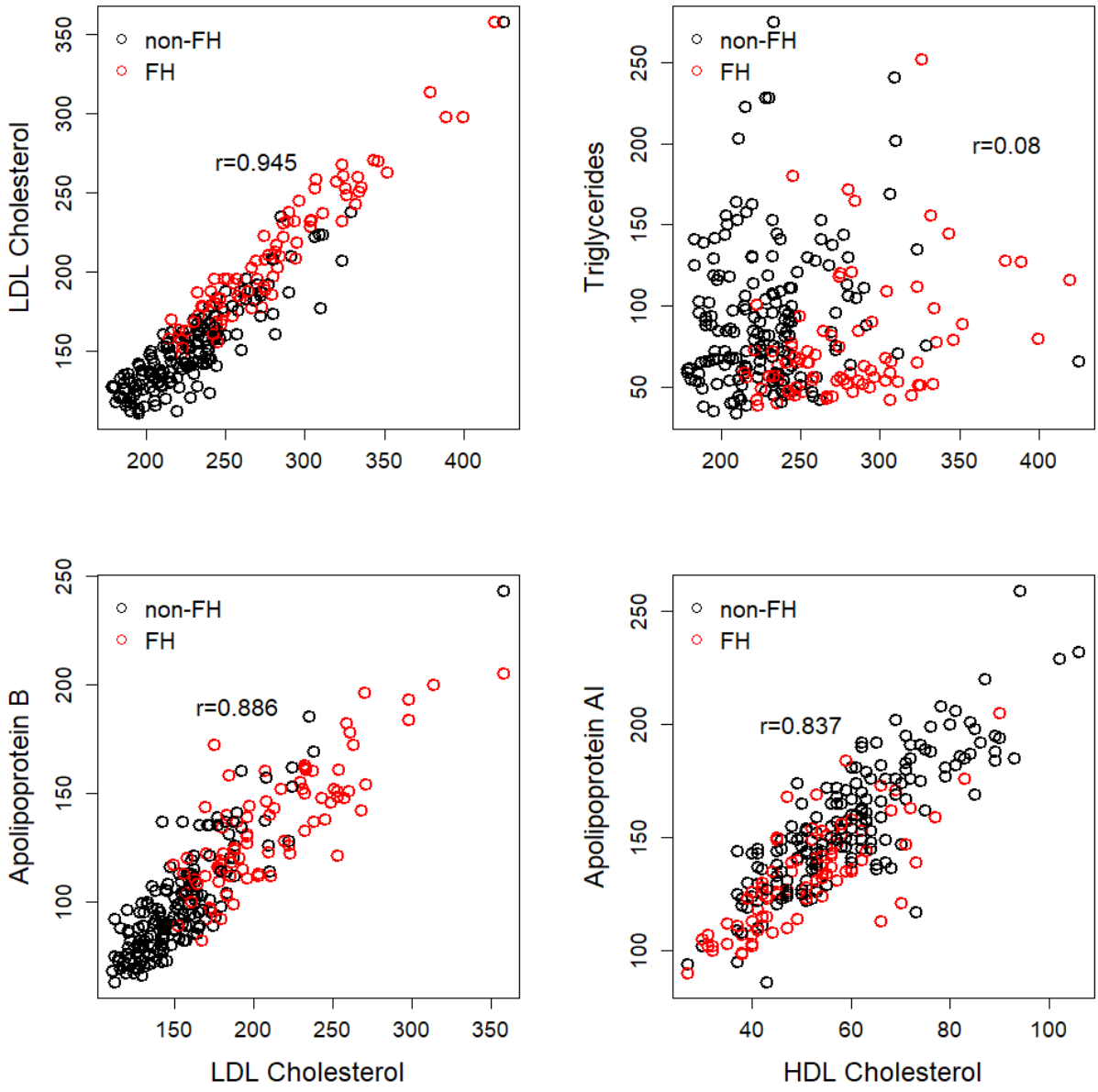


Figure 4: Scatterplots between different pairs of biochemical variables. FH individuals are represented in red, while non-FH cases are represented in black. Pearson correlation coefficients are presented, considering data as a whole.